Electronic Thesis and Dissertation Repository

11-5-2018 2:00 PM

From Accessibility and Exposure to Engagement: A Multi-scalar Approach to Measuring Environmental Determinants of Children's Health Using Geographic Information Systems

Martin Healy, *The University of Western Ontario*

Supervisor: Gilliland, Jason, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Geography
© Martin Healy 2018

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Geographic Information Sciences Commons, and the Human Geography Commons

Recommended Citation

Abstract

A growing body of research suggests that increasing the accessibility to health-related environmental features and increasing exposure to and engagement in outdoor environments leads to positive benefits for the overall health and well-being of children. Additionally, research over the last twenty-five years has documented a decline in the time children spend outdoors. Outdoor activity in children is associated with increased levels of physical fitness, and cognitive well-being. Despite acknowledging this connection, problems occur for researchers when attempting to identify the child's location and to measure whether a child has made use of an accessible health-related facility, or where, when and for how long a child spends time outdoors.

The purpose of this thesis is to measure children's accessibility to, exposure to, and engagement with health-promoting features of their environment. The research on the environment-health link aims to meet two objectives: 1) to quantify the magnitude of positional discrepancies and accessibility misclassification that result from using several commonly-used address proxies; and 2) to examine how *individual-level , household-level*, and *neighbourhood-level* factors are associated with quantity of time children spend outdoors. This will be achieved by employing the use of GPS tracking to objectively quantify the time spent outdoors using a novel machine learning algorithm, and by applying a hexagonal grid to extract built environment measures.

The aim of this study is to identify the impact of positional discrepancies when measuring accessibility by examining misclassification of address proxies to several health-related facilities throughout the City of London and Middlesex County, Ontario, Canada. Positional errors are quantified by multiple neighbourhood types. Findings indicate that the shorter the threshold distance used to measure accessibility between subject population and health-related facility, the higher the proportion of misclassified addresses. Using address proxies based on large aggregated units, such as centroids of census tracts or dissemination areas, can result in vast positional discrepancies, and therefore should be avoided in spatial epidemiologic research.

In an effort to reduce the misclassification, and positional errors, the use of individual portable passive GPS receivers were employed to objectively track the spatial patterns, and quantify the time spent outdoors of children (aged 7 to 13 years) in London, Ontario across multiple neighbourhood types. On the whole, children spent most of their outdoor time during school hours (recess time) and the non-school time outdoors in areas immediately surrounding their home.

From these findings, policymakers, educators, and parents can support children's health by making greater efforts to promote outdoor activities for improved health and quality of life in children. The aim of this thesis is to advance our understanding of the environment and health-link and suggests practical steps for more well-informed decision making by combining novel classification and mapping techniques.

Keywords

Co-Authorship Statement

Each of the manuscripts contained within this dissertation has been published in or prepared for publication in peer-reviewed journals. Chapter 3 and Chapter 4 have been written by Martin Healy with Dr. Jason Gilliland as co-author. In each manuscript, Martin Healy was the principal author and performed all research and analysis, participated in the data collection and database design, and programmed all custom computer programming tools. Dr. Jason Gilliland was involved in the development of the methodological and analytical protocols utilized in each of the two studies. Below is a list of the journal destinations for each of the manuscripts.

Chapter 3: Healy, M. and Gilliland, J. (2012). Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research. Spatial and Spatio-temporal Epidemiology 3: 55-67.

Chapter 4: Healy, M. and Gilliland, J. Objectively measuring children's accessibility, exposure, and engagement in time spent outdoors using passive GPS devices. Prepared for the International Journal of Health Geographics.

Acknowledgments

I want to thank my supervisor, Dr. Jason Gilliland for his steadfast belief in me and his endless encouragement throughout all the stages of my Ph.D. research. I would not have been able to complete this dissertation without his support. I would also like to thank the members of my committee including Dr. Jacek Malczewski, Dr. Micha Pazner, and Dr. Isaac Luginaah and to my examination committee for your careful consideration of this thesis and your constructive comments.

Thanks go out to my second reader Dr. Jinfei Wang who offered me her time to read and make thoughtful comments during a hectic time at the beginning of the academic year. Additional thanks to Lori Johnson and Dr. Jeff Hopkins from the Dept. of Geography who encouraged me to persevere.

A special thanks to Dr. Andrew Clark, with whom I have spent many hours discussing the complexities of the data from the STEAM project. I also wish to thank my colleagues whom I had the pleasure to work closely with over the years including Suzanne Tillman, Kate Schieman, Sandra Kulon, Dr. Janet Loebach, and Kathy Tang.

I thank those HEAL Lab colleagues who have befriended me over the years and from whom I have been inspired to reach my educational goals including Dr. Donald Lafreniere, Dr. Mathew Novak, Dr. Claudia Rangel, Dr. Doug Rivet, and Dr. Richard Sadler. To all past and present members of the HEAL Lab especially Katherine Wilson, and Mohammad El-Bagdady for making me feel part of the team every time I visited the lab. I thank you!

I am grateful for the support of my employer Fanshawe College including my Chair Dr. Dana Morningstar of the School of Design along with my fellow faculty members and I thank my parents and my teachers for instilling in me a love of Geography.

If it were not for Divine Providence, I would not have started my career in GIS which then led me back to Western to upgrade my BA, receive a Masters, and then to write this thesis.

I finally would like to thank my wife Lisa for her love, patience, and encouragement through these many years. I am truly blessed.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1   Introduction

1.1   Research Context

The rise of certain chronic health issues over the past half-century have led researchers and policymakers to place greater emphasis on exploring and identifying potential environmental influences on human population health (Lopez, 2011). Indeed, concerns for the rise in children's health issues, particularly the profound increases in sedentary behaviour, obesity, and mental health problems has recently promoted community planning, and it's product, the built environment, at the forefront of these types of academic studies (Lopez, 2011). Researchers from several academic disciplines, including geography, planning, epidemiology, health promotion, and psychology, have been investigating the role that the built environment has in promoting healthy outdoor behavior (Gilliland, 2010).

Canadian children today, on average, spend less than one hour per day outside (Zorzi & Gagne, 2012) and children between the ages of 8 and 18 years spend an average of six and a half hours a day with electronic media (Roberts et al., 2005). The less time spent outdoors in natural environments has been linked to decreased physical activity (Schaefer et al., 2014; Wheeler et al., 2010), increased rates of obesity (Ansari et al., 2015; Schaefer et al., 2014), and increased rates of myopia (French et al., 2013; Guggenheim et al., 2012; Guo et al., 2013; Rose et al., 2008), sleep disorders, mental health issues (Tillmann et al., 2018), cognitive health issues (Wells, 2000) and nature deficit disorder (Driessnack, 2009; Louv, 2008) in children. Later in life, the accumulation of inactivity raises the odds of a person developing chronic diseases, such as Type-2 diabetes, cancers, and depression (Gilliland, 2010).

Methodological problems abound in the existing built environment and health literature, particularly with respect to how the measurement of accessibility to, exposure to, and engagement with, health-related environmental features (e.g. parks, grocery stores, and recreation centres) are mapped and analyzed in a geographic information system (GIS).

The spatial data used in geographic research are always intrinsically uncertain (Zhang & Goodchild, 2002) and care is required so that the uncertainty does not affect the statistical associations being evaluated. It is often the case that researchers accept that their methods need to include some analysis on the accuracy of the data, but few researchers endeavour to do so. It is commonplace in studies of accessibility and exposure for researchers to use geographic proxies to represent a subject's actual location (e.g. census tracts, dissemination areas, or postal codes). These proxies do not, indeed cannot, accurately represent the physical location of their subjects at all times, and therefore the use of proxies leads to 'distance errors'(Zandbergen, 2007). It is also commonplace that large administrative areal units (e.g., census tracts or county boundaries) are used to assign *neighbourhood-level* attributes, which may introduce additional errors into the research, such as 'accessibility or exposure misclassification', the 'modifiable area unit problem' (MAUP) (Openshaw, 1984), and the 'uncertain geographic context problem' (UGCoP) (Kwan, 2012b). These types of errors can, to some degree, be mitigated by the use of GPS tracking (Cooper et al., 2010; Ellis et al., 2014; Rainham et al., 2008), but other confounders lurk when classifying or binning the GPS data. In particular, errors arise when researchers try to measure time spent outdoors (Cooper et al., 2010; Ellis et al., 2014) due to common weaknesses in how GPS signals are processed. The aforementioned methodological problems will be discussed in further detail in Chapter 2.

Based on the large and growing body of research evidence, the overarching assumption behind this thesis research is that the built environment provides a child with the opportunity to facilitate outdoor activity that will lead to better health and quality of life. It is argued here, however, that serious problems may exist in previous studies which link environment and health based on the mapping of home locations using inappropriate spatial reference data. Additionally, significant exposure misclassification exists when using overly simplistic methods of accessibility (e.g. proximity), which are atemporal, to represent one's interactions with or engagement within an environment, such as time spent outdoors. The studies in this dissertation are woven together through the common goal of improving methodological rigor in the measurement of children's accessibility to, exposure to, and engagement with health-related features of their environment to

ultimately better our understanding of the links between environment and children's health.

It is imperative that researchers identify the extent to which these methodological problems can affect statistical outcomes and to present solutions to these problems through the use of more rigorous methodologies and empirically generated data. In light of the dramatic increase in the time children are spending indoors in sedentary lifestyles and the purported impacts this behavior has on their health, an essential contribution to science and public health would be to develop, test, and validate improved methods for understanding how built environment factors influence childhood health. The primary goal of this thesis, therefore, is to establish more rigourous methods to measure children's accessibility to, exposure to, and engagement in, their outdoor environment.

1.2   Geographic Context

The geographic context of this research is identified in this chapter while a more detailed rationale for choosing the particular study areas will be outlined in Chapters 3 and 4. Chapter 3 is situated in both the City of London (population 350,200) and neighbouring Middlesex County (population 69,024) in Southwestern Ontario, Canada. These two municipalities are ideal study areas for examining the geocoding errors in accessibility studies as they encompass a mix of urban, suburban, small town, and rural agricultural areas (Statistics Canada, 2011). Chapter 4 is set within the city of London, Ontario. London includes an array of built environments ranging from older (pre-WWII), dense urban environments with mixed land uses and grid-like street patterns, as well as newer suburban areas, which are primarily lower density with predominantly residential land uses and curvilinear street patterns. Given London's development patterns and overall built form, the methods and findings in this dissertation are broadly relevant to other mid-sized and smaller Canadian cities.

1.3   Dissertation Organization

This thesis uses a multi-scalar approach to analyze the built environment and the accessibility to, exposure to, and engagement in health-promoting and health-damaging features for children. Although all the research components share a common theme of understanding the role of the built environment, the themes are sufficiently different to merit an integrated-article format for this dissertation.

Chapter 2 reviews some of the mounting body of evidence on how children's interactions with the outdoors can influence their physical, mental, and cognitive health. The review illustrates the growing consensus of children's health researchers on the benefits of 'being outdoors'. This review shows that there is strong evidence to support the hypothesized relationship between children's interactions with the outdoors and their health, and thereby justifies the need for the quantitative methods and research presented here. The first part of the chapter will give a brief overview of the literature on environmental influences on children's health and well-being, with a specific focus on the benefits of being outdoors. The second part will focus on the issue of 'uncertainty' in geographic analyses, with particular consideration of the implications of uncertainty in spatial data, GPS tracks, data classification and spatial analyses when mapping human subjects and the built environment.

The purpose of the *neighbourhood-level* study in Chapter 3 is primarily to examine the misclassification of accessibility when associating a sample unit with a proxy location for a child's home address (address proxy). The study quantifies the magnitude of positional discrepancies and accessibility misclassification that result from using several commonly-used address proxies in public health research. The impact of these positional discrepancies on spatial epidemiology is illustrated by examining misclassification of accessibility to several health-related facilities throughout the City of London and Middlesex County, Ontario, Canada.

The home location proxies will be examined to identify the misclassification of accessibility to several health-related facilities, as well as to quantify the shortest path positional errors of the proxies across multiple neighborhood types (rural, small town,

suburban, and urban), in order to reveal the utility of each address proxy for each neighbourhood type and inform future geographic health studies. The research objectives for this study include answering the following questions:

1. When choosing an address proxy what should a researcher expect regarding positional errors when measuring shortest path distance from that proxy to health-related facilities (Junk Food, Grocery Stores, Schools, Recreational Facilities, and Hospitals) by neighbourhood type (rural, small town, suburban, and urban)?

2. When creating network distance buffers originating from the centroid of each commonly used address proxy (e.g. Census Tracts, Dissemination Areas, Geocoded Address) what is the percentage that the health-related facilities contained within the buffer are misclassified by neighbourhood type?

Chapter 4, is divided into two separate but complimentary studies to ultimately measure children's exposure to, and engagement in their outdoor environment. The study addresses the insufficient methods of identifying outdoor activity in children who wear passive GPS receivers, by proposing a novel protocol using a combination of 1 second epochs of GPS data collection, a proven GIS kernel based method of identifying routes and stops, distance measurements to buildings, a random forest model, and the use of a hexagon tessellated surface. The first part of Chapter 4 will focus on the methodology to verify and classify GPS coordinates as occurring indoors or outdoors. The methodology can be employed in future studies where subjects are tracked with GPS receivers without the need to use a particular brand of receiver. The study suggests combining three methods to catalogue, classify, and bin GPS tracks and then tests the methodology on a large GPS dataset generated from a sample of children. Once classified and binned, the GPS data will be used to measure the time children spend outdoors. The second part of the study will make use of the outdoor GPS tracks to measure the exposure to and engagement of the child participants in the built environment. The GPS tracks, coinciding with a hexagonal tessellation surface of built environment features, are used to measure

exposure to those features, while the elapsed time spent in each hex bin acts as a proxy for engagement with those same features.

The research objectives and questions for this study in Chapter 4 include:

1. Can machine learning algorithms be used to identify whether the GPS was indoors or outdoors during its operation?

2. Does seasonality, the *neighbourhood-level* built environment, *household-level* socio-economic status, and *individual-level* age and sex influence the amount of time children spend outdoors on weekdays and weekends? How do the findings contrast with previous research of children's time spent outdoors?

3. Where do children spend time the outdoors? For those outdoor spaces children use, how long are they being used? How do the measurements of exposure contrast to the measurements of engagement?

Chapter 5 summarizes and discusses the key research findings, outlines the limitations of the work, offers conclusions, and proposes next steps for future research.

1.4   Conceptual and Methodological Framework

The factors that influence an individual's health are complex and cannot be fully explained using the biomedical approach (Engel, 1977; Hill, 1965). Engel (1977) challenged the status quo to suggest that biological factors act in combination with an individual's experiences, and their social and psychological factors to determine an individual's susceptibility or resiliency to disease. Bronfenbrenner's (1977) ecological model describes the context in which children develop. He imagined spheres of influence beginning with the individual child as the centre, surrounded by and interacting with the immediate influences of home, school, and neighbourhood in a sphere called the microsystem (See Figure 1.1). He argued that children's development could not be thought of as being independent from the multi-leveled social, material and cultural context (Mesosystem, Exosystem, and Macrosystems) in which a child's development

takes place. Bronfenbrenner's model was developed within his own discipline of psychiatry, and he did not focus on the role that geographic phenomena have on the individual in terms of accessibility to, exposure of, and engagement at health-promoting and health-demoting
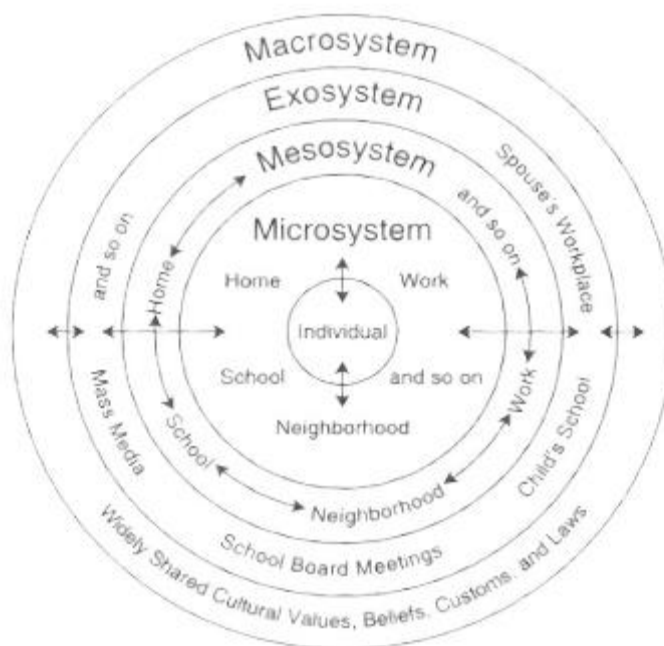
Figure 1.1: Bronfenbrenner's (1977) ecological model

built environment locations.   The socio-ecological model of health behaviour conceptualizes that individual health outcomes are not only a result from individual behaviour but, moreover, from a series of relationships between individuals and their environments (e.g., neighbourhood, work, school) which then informs that child's behaviour (Sallis et al., 2006). The socio-ecological influences on individuals can then lead to modifications in their health behaviour and status, both positively and negatively. The socio-ecological model is well suited to help describe why an individual's geographical, environmental and social context can act as a hindrance to, and a facilitator of, health-related behaviours (Sallis et al., 2006; Stokols, 1992). The model was further refined by Sallis and colleagues (2006) to include perception of environmental factors such as safety and accessibility. Ecological models are widely used and seem well equipped to conceptualize the complex relationship between children, their health-related

behaviours, and their environment. Therefore, this thesis will employ the socio-ecological model of human behaviour approach to understand the multi-layered factors that influence children's health.

The modified Bronfenbrenner (1979) socio-ecological model (see Figure 1.2) includes his idea of a time sphere called the chronosystem. Additional modifications were imposed on the model in terms of the concrete ways in which this thesis will operationalize the individual, microsystem, mesosystem, and chronosystem interactions to measure children's accessibility to, exposure to, and engagement with health-promoting and health-demoting features of their environment. Firstly, the microsystem and mesosystem spheres will be referred to in this thesis as the *household-level* and *neighbourhood-level* environments, while chronosystem is referered to a the *temporal-level*. Secondly, the model in Figure 1.2, indicates the types of variables used in this thsis that are theorized to play some role in the environmental determinants of children's health.

The purpose of this thesis is to measure children's accessibility to, exposure to, and engagement with health-promoting features of their environment. The research on the environment-health link aims to meet two objectives: 1) to quantify the magnitude of positional discrepancies and accessibility misclassification that result from using several commonly-used address proxies; and 2) to examine how *individual-level , household-level*, and *neighbourhood-level* factors are associated with quantity of time children spend outdoors. This will be achieved by employing the use of GPS tracking to objectively quantify the time spent outdoors using a novel machine learning algorithm, and by applying a hexagonal grid to extract built environment measures.

A geographic study that employs the socio-ecological approach is logical in that an individual child experiences and engages with their environment at specific times, locations and places. The socio-ecological model in combination with geographic analysis, therefore, is particularly well-suited for a studying children's accessibility to, exposure to, and engagement in their environment which in turn plays a crucial role in

their healthy development. This study, therefore uses a positivist spatial quantitative approach to practically measure, classify, categorize and map children and their neighbourhoods.
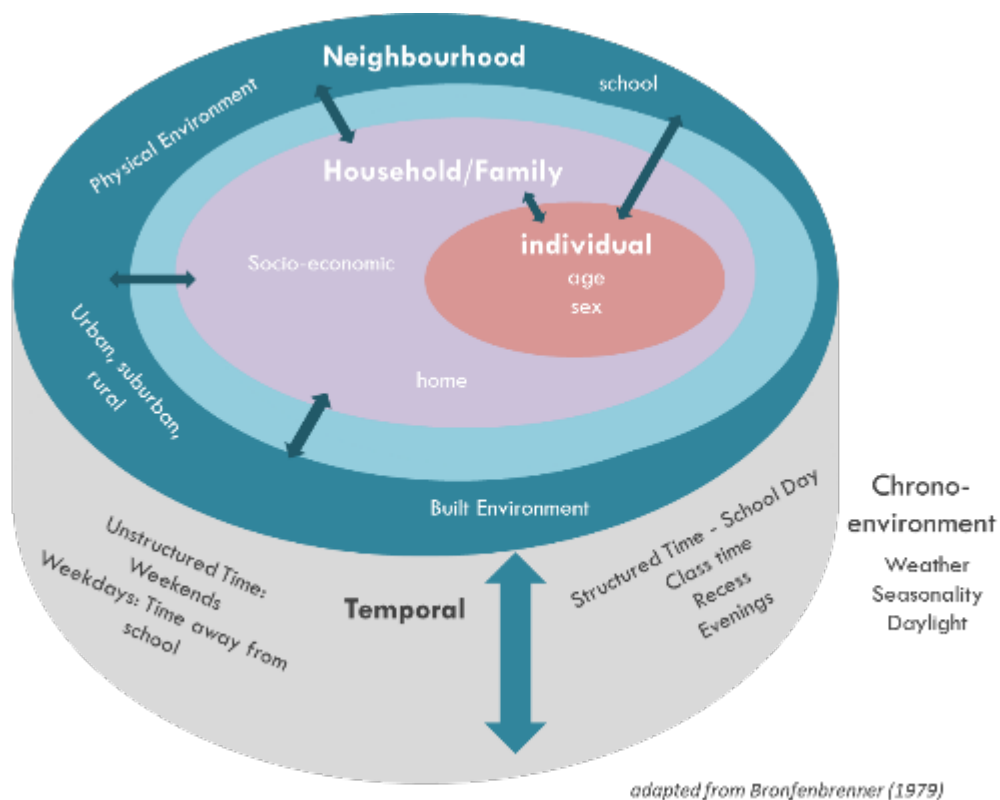
Figure 1.2: Spatio-temporal model of child's local environmental accessibility, exposure, and engagement (Bronfenbrenner, 1979)

.

1.5 References

Ansari, A., Pettit, K., & Gershoff, E. (2015). Combating obesity in head start: Outdoor play and change in children's body mass index. *Journal of Developmental and Behavioral Pediatrics, 36*, 605-612.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513-531.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design.* Cambridge, Massachusetts: Harvard University Press.

Cooper, A. R., Page, A. S., Wheeler, B. W., Hillsdon, M., Griew, P., & Jago, R. (2010). Patterns of GPS measured time outdoors after school and objective physical activity in English children: the PEACH project. *Int J Behav Nutr Phys Act, 7*, 31. doi:10.1186/1479-5868-7-31

Driessnack, M. (2009). Children and nature-deficit disorder. *Journal for Specialists in Pediatric Nursing, 14*(1), 73-75. doi:10.1111/j.1744-6155.2009.00180.x

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Front Public Health, 2*, 36. doi:10.3389/fpubh.2014.00036

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science, 196*(4286), 129-136.

French, A. N., Ashby, R. S., Morgan, I. G., & Rose, K. A. (2013). Time outdoors and the prevention of myopia. *Experimental Eye Research*. doi:10.1016/j.exer.2013.04.018

Gilliland, J. (2010). The Built environment and obesity: trimming waistlines through neighbourhood design. . In Bunting, Filion, & Walker (Eds.), *Canadian cities in transition.* (4th ed., pp. 391–410): Oxford Univ Press.

Guggenheim, J. A., Northstone, K., McMahon, G., Ness, A. R., Deere, K., Mattocks, C., . . . Williams, C. (2012). Time outdoors and physical activity as predictors of incident myopia in childhood: a prospective cohort study. *Investigative Ophthalmology and Visual Science, 53*, 2856-2865. doi:10.1167/iovs.11-9091

Guo, Y., Liu, L. J., Xu, L., Lv, Y. Y., Tang, P., Feng, Y., . . . Jonas, J. B. (2013). Outdoor activity and myopia among primary students in rural and urban regions of Beijing. *Ophthalmology*. doi:10.1016/j.ophtha.2012.07.086

Hill, B. A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine, 58*, 295-300.

Kwan, M. (2012b). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers, 102*(5).

Lopez, R. (2011). *The Built Environment and Public Health*. New York, NY,: John Wiley & Sons.

Louv, R. (2008). *Last child in the woods: saving our children from nature-deficit disorder*. Chapel Hill, N.C.: Algonquin Books of Chapel Hill.

Openshaw, S. (1984). *The modifiable areal unit problem.* Norwich: Geobooks.

Rainham, D., Krewski, D., McDowell, I., Sawada, M., & Liekens, B. (2008). Development of a wearable global positioning system for place and health research. *Int J Health Geogr, 7*, 59. doi:10.1186/1476-072X-7-59

Roberts, D. F., Foehr, U., & Rideout, V. (2005). *Generation M: Media in the lives of 8 to 18 year olds*. Retrieved from Menlo Park, CA: http://www.kff.org/entmedia/entmedia030905pkg.cfm

Rose, K. A., Morgan, I. G., Ip, J., Kifley, A., Huynh, S., Smith, W., & Mitchell, P. (2008). Outdoor Activity Reduces the Prevalence of Myopia in Children. *Ophthalmology*. doi:10.1016/j.ophtha.2007.12.019

Sallis, J. F., Cervero, R. B., Ascher, W., Henderson, K. A., Kraft, M. K., & Kerr, J. (2006). An ecological approach to creating active living communities. *Annu Rev Public Health, 27*, 297-322. doi:10.1146/annurev.publhealth.27.021405.102100

Schaefer, L., Plotnikoff, R. C., Majumdar, S. R., Mollard, R., Woo, M., Sadman, R., . . . McGavock, J. (2014). Outdoor time is associated with physical activity, sedentary time, and cardiorespiratory fitness in youth. *J Pediatr, 165*, 516-521. doi:10.1016/j.jpeds.2014.05.029

Stokols, D. (1992). Establishing and maintaining healthy environments: toward a social ecology of health promotion. *American Psychologist, 47*(1), 6-22.

Tillmann, S., Tobin, D., Avison, W., & Gilliland, J. (2018). Mental health benefits of interactions with nature in children and teenagers: a systematic review. *J Epidemiol Community Health*. doi:10.1136/jech-2018-210436

Wells, N. M. (2000). Effects of Greenness on Children's Cognitive Functioning. *Environment and Behavior, 32*, 775-795.

Wheeler, B. W., Cooper, A. R., Page, A. S., & Jago, R. (2010). Greenspace and children's physical activity: a GPS/GIS analysis of the PEACH project. *Prev Med, 51*(2), 148-152. doi:10.1016/j.ypmed.2010.06.001

Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health, 7*, 37. doi:10.1186/1471-2458-7-37

Zhang, J., & Goodchild, M. (2002). *Uncertainty in Geographical Information.* London: Taylor & Francis.

Zorzi, R., & Gagne, M. (2012). *Youth Engagement with Nature and the Outdoors: A Summary of Survey Findings*. Retrieved from https://davidsuzuki.org/wp-content/uploads/2012/09/youth-engagement-nature-outdoors.pdf

Chapter 2

2   Literature Review

The literature review is divided into two parts. The first part will give a brief overview of the literature on environmental influences on children's health and well-being, with a specific focus on the benefits of 'being outdoors'. The second part will focus on the issue of 'uncertainty' in geographic analyses, with particular consideration of the implications of uncertainty in spatial data, data classification and spatial analyses when mapping human subjects and the built environment.

2.1   Benefits of Outdoor Accessibility, Exposure and Engagement for Children

Recently, there has been widespread public attention and a surge in academic literature published on the health benefits of spending time outdoors, especially on the child population (Tremblay et al., 2015). Research shows that spending time outdoors can positively impact children's physical activity (Cleland et al., 2008), mental health, well-being, social health, and cognitive development. Dramatic increases in sedentary behaviour and time spent using electronic devices is a concern of many parents, practitioners, and researchers, which helps support further investigation into the relationship between time spent outdoors and a variety of children's health outcomes.

A body of the literature identified in a recent systematic review assessed the effect of outdoor time on children's physical activity, sedentary behaviour, and physical fitness (Gray et al., 2015). Some studies included in this systematic review agreed that there were overall positive effects of outdoor time on physical activity, sedentary behaviour, and cardiorespiratory fitness (Gray et al., 2015). Each study assessing physical activity found higher levels outdoors compared with indoors. Distinguishing activities based on whether it is happening indoors or outdoors is vital as studies have shown that children are more active in outdoor environments (Raustorp et al., 2012).

Previous research has identified the mental health benefits of interactions with nature outdoors into three types: accessibility, exposure, and engagement, as documented by

Tillmann et al. (2018). Total outdoor exposure is defined as direct contact with outdoor environmental features and engagement outdoors is defined by the total time spent outdoors. Tillmann and colleagues (2018) state that some of these studies using measures of accessibility such as residential proximity to outdoor greenspace have critical weaknesses when assessing children's interaction with their natural environments. The most fundamental problem is that there is no proof that children are making use of these spaces. Research should therefore be accounting for the individual choices made by children when discussing interactions with particular environments. Exposure to and engagement with the outdoors, therefore, might give us more accurate representation of the actual time children spend at particular locations.

One of the significant barriers of assessing children's time spent in outdoor environments is the inability to precisely determine whether a child's GPS tracks are indoors or outdoors. Previous research has used time blocks (e.g., a school schedule) to determine whether a child is most likely indoors or outdoors (Loebach & Gilliland, 2014). However, this has limitations in that it assumes that all participants are in the same environment based on a school schedule. Some studies have supplemented time blocks with self- or parent-reported activity diaries detailing where children are; however, this again makes some assumptions based on time blocks included in the diary and creates an opportunity for inaccurate reporting by children (Loebach & Gilliland, 2014). Self or parent reports have also been used to classify use or time spent in specific spaces which again leaves room for inaccurate reporting as well as not being an accurate representation of every space a child interacts with on a daily basis (Amoly et al., 2014; Faber Taylor & Kuo, 2011; Flouri et al., 2014; McCracken et al., 2016). Being able to accurately determine whether a single GPS point is indoors or outdoors is crucial for more accurately accounting for a child's activity choices.

2.2   Geographic Data and Uncertainty

Uncertainty is a form of ignorance that Thrift (1985) argues has five forms which include; a lack understanding; not knowing the unknown; issues left undiscussed or deliberately hidden; and that which is distorted. Uncertainty abounds in every research

study, especially when including geographic information systems as a tool for analysis. Geographic information is a digital representation of an abstract view of reality (David et al., 1996). Therefore, it is impossible to perform error-free spatial analysis, and it is each researcher's responsibility to identify and mitigate these errors to such a degree that they do not interfere with the conclusions derived from that analysis (S. A. Fotheringham, 1989). Over the succeeding years, considerable effort has been spent by researchers trying to remove uncertainty from GIS analysis. Couclelis (2003) contends that there will always be uncertainty in any scientific study, not just those studies using GIS, and she argues for the acceptance that some error and uncertainty will never entirely be removed and should be considered part of the process of the exploration in science.

It is understood that the data and the methods used in this thesis will be imperfect and will, in turn, generate results that will be somewhat uncertain. One of the purposes of this research is to identify where this uncertainty lays, regarding the GIS data used and in the GIS methods proposed, and suggest ways to mitigate some of these uncertainties. There are two main discussions presented in this part of the literature review. The first discussion will focus on the types and magnitude of spatial data errors inherent in GIS and GPS data. It is from the spatial data errors that some uncertainty will always be introduced; furthermore, a plan to identify and mitigate these errors needs to be explored. The second discussion found later in this section includes a short critique of the commonly used GIS methods used to associate neighbourhood built environment variables with children's health outcomes and the implications of the modifiable areal unit problem and uncertain geographic context problem on this research.

2.2.1   Uncertainty with Spatial Data

All spatial data are intrinsically uncertain in a world that is infinitely complex (Zhang & Goodchild, 2002). Measurements of weather conditions, air and water pressure, prevailing winds, pollution levels in the air and water, population densities, income levels, accessibility, and the movement of people all vary continuously. A large part of the uncertainty generated with GIS spatial analysis originates from the quality of the data itself. The quality of GIS datasets can divide into two separate aspects, precision and accuracy. Precision refers to the resolution, or amount of detail in the GIS data regarding

positional, thematic, and temporal dimensions. Accuracy is defined as the 'inverse of error'. Researchers must be aware of the difference between what is spatially and thematically encoded in a data set and what should be encoded in that data set (Albrecht, 2007); in other words, researchers must realize that there will always be something missing in a data set. Accuracy, therefore, is a relative term rather than an absolute one. Depending on the purpose of a data set, a researcher might have higher or lower expectations of accuracy depending on the purpose. The precision of the data plays a part in this as well; if a data set is of low spatial precision, then the researcher would, in turn, have lower expectations of accuracy whenever using this data. Temporal accuracy describes the difference in the recorded time of an event to the actual time of the event, while thematic accuracy describes the concurrence of what is encoded in an attribute table and what should be encoded.

When modelling the 'real-world' in a GIS, the complexities of geographic phenomenon will be lessened through map generalization. The goal of digital map generalization is to maintain the graphic detail of the map features while at the same time simplifying them so that geographic features of shape, size, and position are faithful to what they represent at the map scale for which they will be used (Buttenfield, 1991). The act of map generalization reduces the precision of the real-world feature modelled, thus affecting the expectations of accuracy. The absolute positional accuracy of a spatial feature is calculated by measuring the difference between the recorded location in a GIS dataset, and the feature's true location. Relative positional accuracy of a spatial feature is calculated by measuring the difference between the recorded location in a GIS dataset, and a location of a corresponding feature in another GIS data set. Positional (absolute or relative) errors are the differences between these matching features and their coordinate locations. For point features, the error can be defined in x, y, and z dimensions and the metrics describing the error can use simple descriptive statistics. For lines and areas, more complex methods for generating the accuracy metrics including using buffers (Goodchild & Hunter, 1997; Tviete & Langaas, 1999) and stochastic simulation techniques (Leung & Yan, 1998; Shi & Liu, 2000; Zhang & Kirby, 2000) are required. Other terms for positional error are 'displacement' and 'distortion' (Zhang & Kirby, 2000). So it could be stated that the presence of absolute or relative spatial distortions

will affect distance, and density measurements in any GIS study, and will, therefore, introduce spatial uncertainty.

In a GIS there is both the spatial features and tabular data. The tabular data stores the quantitative and qualitative information about each of the geographic features. The tabular data can and often does suffer errors which can occur at the database design or data modelling stages, and at the data entry phase. The concept of thematic accuracy, therefore, is the accuracy of the attribute values encoded in a GIS database. The metrics used to describe accuracy depend on the measurement scale of the data. Quantitative data accuracy can be measured and errors identified by using simple descriptive statistics such as standard deviation, minimum, maximum, and mean. The qualitative data can be assessed with a classification error matrix by using cross-tabulation, at a series of sample locations, to match what feature is present against the feature that was encoded. If the entire data set were assessed using the cross tabulation method, it would be possible to attach an accuracy attribute on individual features.

There is some disagreement in the literature regarding temporal accuracy. Some consider temporal accuracy to be a function of the latency period between a change of a spatial feature in the 'real-world' and seeing that change reflected in a GIS database (Aalders, 2002; Goldberg, 2008). The other approach to modelling temporal accuracy concerns whether or not the GIS data set has a time dimension connected to the spatial information resulting in the fourth dimension being stored (x,y,z,t) (Aalders, 2002; Thierry et al., 2013). When assessing temporal accuracy, it is necessary to investigate the temporal coordinate in relation to the other three coordinates to then identify correlations between space and time to identify anomalies in the time that was encoded.

Before performing any spatial analysis, an assessment of the quality of the data to be used in the analysis must be undertaken. Data quality represents, for the researcher, the suitability for the use of the data for any particular application. There is no 'one approach' to assess the suitability of GIS data. The researcher must apply a strategy that takes into account the type of analysis to be performed, the nature of the results, and in what manner the results will be transformed from data into knowledge (Albrecht, 2007;

Couclelis, 2003). For the last idea, a distinction should be made between measurements of internal and external data quality. Internal data quality measures the specifications of how the data were collected and processed. External data quality suggests to researchers on how well any particular data will fit their particular application (Aalders, 2002).

2.2.2   Uncertainty with GPS Data

All GPS units execute a three-dimensional trilateration calculation in the generation of a single coordinate. In order for a precise coordinate to be calculated, the GPS unit requires 4+ GPS satellite radio signals to calculate the distance ranges from the satellite to the unit itself. Furthermore, it is critical that the satellites utilized in the distance range calculation should be well distributed in the sky to reduce dilution errors which manifest as positional errors in the creation of the GPS point coordinate. It is expected that positional errors will occur when a participant wearing a GPS unit enters/exits or remains inside a blocking structure such as a building or dense tree canopy, thus blocking, in whole or in part, the sky. When part of the sky is blocked from view, a GPS unit will use the available satellites (space vehicles) from the GPS constellation that are in its line-of-sight to generate a coordinate. If the GPS unit has line of sight to only a portion of the sky then the accuracy of the GPS coordinate will be compromised. Additionally, for a short time, while the unit is initially started there will be coordinates with larger spatial errors as the GPS unit begins acquiring the GPS satellite signals.

The GPS unit, if programmed to do so, will store a series of quality metrics for each coordinate generated. These metrics follow the coding standard developed by the National Marine Electronics Association (NMEA). Each coordinate has a corresponding NMEA quality sentence data structure. The quality metrics can include, the PDOP (Positional Dilution of Precision), HDOP (Horizontal Dilution of Precision), SNR (Signal to Noise Ratio), NSAT (Number of satellites used to calculate the coordinate), and two qualitative accuracy values '2DGPS' (2-dimensional bias remedied) and '3DGPS' (3-dimensional bias remedied) differential accuracy. The differential accuracy refers to the successful inclusion of the Wide Area Augmentation System Satellite signal in the removal atmospheric interference bias found with the GPS satellite radio signal.

For most applications using GPS technology, the wearer or operator of the GPS device decides when and where to capture a coordinate. Depending on the accuracy available at the location, the operator might choose not to generate a coordinate due to the large error. In passive GPS data acquisition, the GPS is preset to capture GPS coordinates at a set epoch which leads to an enormous number of points that need to be post-processed in some way so that those points with larger errors can be remedied or filtered.

Several researchers in spatial health studies have developed methods to filter, categorize and remedy the erroneous points generated with passive GPS data collection (Patrick et al., 2008; Rainham et al., 2012; Rainham et al., 2008; Thierry et al., 2013). Included in this effort is the Personal Activity and Location Measurement System (PALMS) (Patrick et al., 2008) which filters and smooths the GPS data points by removing invalid coordinates and reducing data volume. PALMS filters data by removing GPS points that indicate excessive speed; that have large changes in elevation or with very small changes in distance between consecutive points; and PALMS reduces the scatter caused by interference from buildings (Kerr, Norman, et al., 2012; Patrick et al., 2008) by employing the NMEA GSV sentence SNR (Signal to Noise Ratio) metric, by doing so the PALMS is limited to only the Qstarz brand of GPS devices. Rainham et al. (2012) created a GIS software tool called the GeoActivity Processor which uses a predefined set of decision rules including known times when the participant was at a geographic anchor area (e.g. home, school, work) and leveraged an assortment of spatial data layers and self-reported diary entries in the decision rules. The GPS points are then grouped into these anchor areas. Thierry et al. (2013), developed a tool called the "Activity place detection algorithm for GPS data" (*SphereLab Tool*) which uses a kernel density approach to filter the GPS points to identify places where the participant stopped for some defined duration. Other spatial health researchers have also employed a variety of classification methods to help filter the raw GPS points while others outside the discipline have made progress with the use of 'big-data analytic approaches' (Brusilovskiy et al., 2016; Kim et al., 2012; Meijles et al., 2014; Wan & Lin, 2013).

2.2.1 Indoor and Outdoor GPS Data Classification

It is unfortunate that some researchers have not classified or filtered their GPS generated points entirely to identify whether a GPS point was generated in the precise locations, while others just visually inspected the spatial errors, and manually removed the apparent errors proceeded with their analysis (Burgi et al., 2016; Elgethun et al., 2002; Elgethun et al., 2007; Maddison et al., 2010; Quigg et al., 2010). Cooper et al. (2010) , Wheeler et al. (2010), and Pearce et al. (2014) while measuring the time children spend outdoors using the Personal and Environmental Determinants of Children's Health (PEACH) project protocol, chose not to classify the GPS points as being generated indoors or outdoors, but instead simply relied on the GPS to 'cut-out' when a child entered a building indicating indoors. They categorized all GPS points recorded as being outdoor time and matched that time (10-second epochs) with a continually running accelerometer (10-second epochs) as 'physical activity outdoors' and any unmatched accelerometer data (no GPS record) as 'physical activity indoors'. These researchers justified their reasoning by using a GPS receiver that did not record positional data when inside a building. By contrast, Kim et al. (2012) tested a GPS receiver that continually generated points regardless of the unit being indoors and outdoors. They employed the use of the GPS point quality metrics of speed and number of satellites (NSAT) and distance from home. They employed a field technician to follow a highly scripted set of indoor and outdoor activities and the locations of these activities while keeping a record of all movements by the second in a diary. They classified their GPS data into four microenvironments (residential indoors, other indoors, transit, and walking outdoors). GPS points were classified as 'indoors' of the time when the NSAT metric was less than 9 and coded as 'residential indoors' 97% of the time when these 'indoor' GPS points were within 40m of home, while the remaining GPS points were classified as outdoor locations.

Researchers who employ a GPS receiver that supports the NMEA GSV protocol have employed the personal activity and location measurement system (PALMS) data filter which classifies GPS points as occurring indoor or outdoor using the signal to noise ratio (SNR) metric (J. Carlson et al., 2015; Ellis et al., 2014; Gell et al., 2015; Kerr et al., 2011; Kerr, Marshall, et al., 2012; Klinker, Schipperijn, Christian, et al., 2014; Klinker,

Schipperijn, Kerr, et al., 2014; Klinker et al., 2015; Lam et al., 2013; Tandon et al., 2013). Specifically, any GPS points with an SNR < 250 on a 0-450 scale are classified as being generated indoors. A strong signal (>=250) suggests the wearer of the GPS is likely to be outdoors where there is less interference from buildings and natural canopies. Presently the PALMS is limited only to the Qstarz brand GPS device to classify tracks as indoor vs. outdoor ("Personal Activity Location Measurement System User Guide," 2011). There have been few studies measuring the validity of the Signal to Noise ratio cut-off method used by PALMS to classify outdoor time, except for Lam et al. (2013) who employed a self-activating camera in combination with a passive GPS monitor to measure time spent outdoors by adults. They found while using 15-second epochs for each GPS point that 81% of the GPS points classified as indoors by PALMS were correct. Tandon et al. (2013), while studying the outdoor activities of pre-school aged children found an 82% match between PALMS coded outdoor activity and an objectively generated measurement of outdoor activity. Klinker, Schipperijn, Kerr, et al. (2014) found while studying the outdoor weekday patterns among school children in Denmark that PALMS overestimated the time that children spent outdoors. J. Carlson et al. (2015) found an 88% predictive value when using PALMS to classify modes of travel at the minute epochs. Most recently, Pearce et al. (2018) while using the Qstarz brand of GPS at epochs of 10 seconds performed their own signal-to-noise classification and identified SLR>=212 as the low cut-off for outdoors, less by 38 points than the PALMS threshold. They suggested the lower cutoff was related to the built form of the neighbourhood setting from their study. If Pearce et al. (2018) are correct and the neighbourhood setting plays such a large effect on the signal-to-noise ratio cutoff then this throws some shadow on the efficacy of studies relying on the standard cutoff (SLR>=250) of the PALMS protocol.

The methods employed by researchers to classify the spatial context of GPS points outside the of the PALMS data filter and a single branded GPS are varied. Increasingly, researchers have been employing spatiotemporal data mining algorithms (Brusilovskiy et al., 2016), Classification and Regression Tree models (Meijles et al., 2014), and kernel density calculations (Han et al., 2013; Kestens et al., 2016; Thierry et al., 2013) to group, filter, and classify their GPS point clouds. Ellis et al. (2014) tested a small set of GPS

point with a naive Bayesian classifier and the random forest model to identify active travel trips. They found that the random forest model classification achieved the best results with an 89.8% cross-validation accuracy while the naive Bayesian classifier had an overall accuracy of 74.2%.

The study by Wu et al. (2011), where the inspiration originated for the use of the random forest model classifier in Chapter 4 of this thesis, suggests that the random forest model for classification could be used effectively if an accurate and large enough training sample could be secured. These researchers compared two automated approaches to classify the GPS points in four ways; as indoor, in-vehicle travel, outdoor static and outdoor walking. They used GPS data from two separate participant studies and found that a rules based approach performed slightly better than the Random Forest for GPS point classification. They suggested that the random forest results suffered from a small and compromised training sample stemming from the flaws in the way the GPS data was post-processed which used a combination of areal imagery of the study area, daily activity diaries and memory recall.

In a recent systematic review of measurement of time spent outdoors in child myopia research, J. Wang et al. (2018) highlighted on the research to date using GPS to differentiate between indoor and outdoor location. The research studies mentioned in their systematic review are the same as in this literature review, except for the most recent of studies, and in their review they report only on the accuracy of the PALMS protocol. They discussed that some studies compromised their ability to differentiate indoor and outdoor because of the use of larger epochs of data collection (e.g. 15 second) which do not capture subtle movements. They, also identified the role that a combination of geographical information system (GIS), diary/questionnaire, and accelerometers have had to improve the accuracy, but they also suggest that the accuracy of GPS devices in general requires improvement. They report that more research is required to improve methods classifying the GPS points (indoor/outdoor) so to increase the validity and accuracy of this type of data.

In Chapter 4 an improved method of differentiating between indoor and outdoor using GPS will be proposed in response to the uncertainty detailed in this section of the literature review. The dependence on a single brand of GPS receiver will be overcome, a commonly used NMEA sentence will be employed, and the issues inherent with small unreliable training samples will be remedied.

2.2.3   Uncertainty with the GIS Methods

Recalling the five forms of not knowing (Thrift, 1985) and that the GIS is a tool used for the combination of geographic information, and consequently, for the production of applied geographic knowledge (Couclelis, 2003), it is imperative that GIS practitioners do not impart additional uncertainty by using GIS methods without careful consideration beforehand. In this part of the literature review, the commonly used methods of geocoding and of spatial aggregation used to associate built environment variables with children's health outcomes are discussed.

Geocoding is the process of converting pseudo-spatial tabular data of street addresses to map coordinates. The process is widely used in environmental and health studies to locate subjects and built environment indicators (Brownson et al., 2009). Anselin (2006) contends that the results from the geocoding are rife with uncertainties. Much research has been conducted with problems with the match rate due to inaccurate address information (Gilboa et al., 2006; Goldberg et al., 2008; Henry & Boscoe, 2008; Lovasi et al., 2007; Mazumdar et al., 2008; Rutt & Coleman, 2005; Whitsel et al., 2006; Zhan et al., 2006; Zimmerman, 2008; Zimmerman et al., 2008; Zimmerman & Li, 2010) and with the defects in the spatial database containing street locations (Hay et al., 2009; Hong & Vonderohe, 2014; Jacquez & Rommel, 2009; Lovasi et al., 2007; Schootman et al., 2007; Strickland et al., 2007; Whitsel et al., 2006; Zandbergen, 2007; Zandbergen & Green, 2007; Zimmerman & Li, 2010; Zinszer et al., 2010). Best geocoding practices are outlined by Goldberg (2008), and many studies account for these types of errors, but in a way that might not be ideal. Anselin (2006) notes that geocoding errors, both the match rates and positional errors tend to be found in newly developed suburban areas with these errors causing biased health outcome metrics in these areas.

There are various ways to identify a subject's neighbourhood so that built environment variables can be assigned and some association reached regarding health outcomes of the population in these neighbourhoods. Some researchers have used census boundaries to delineate the neighbourhood (Chuang et al., 2005; Goovaerts, 2009; Larsen & Gilliland, 2008; J. Pearce et al., 2006; M. C. Wang et al., 2007) and some have used municipal planning districts (J. Gilliland et al., 2006), while other use neighbourhoods as a function of walking distances from the home or school (Ball et al., 2007; C. Carlson et al., 2012; Frank et al., 2004; J. A. Gilliland et al., 2012; Kelly et al., 2007; Lee et al., 2009; Leslie et al., 2005; Li et al., 2005; L. N. Oliver et al., 2007; Panter et al., 2010; Robitaille & Herjean, 2008). Other studies use the 'nearest neighbour' and kernel-based approaches such as Geographic Weighted Regression to generate natural neighbourhoods (Bjork et al., 2008; Clark & Scott, 2014; A. Fotheringham & Wong, 1991; Goovaerts, 2009; Li et al., 2005; Maroko et al., 2009; Swift et al., 2013; Tandon et al., 2015; Webster et al., 2006). All of these methods of neighbourhood delineation impose uncertainty and are prone to the ecological fallacy. Each method aggregates geographic features, and set boundaries or limits which, in turn, impose the modifiable areal unit problem (MUAP) with both effects; zonal and scale (Openshaw, 1984) and impose the boundary problem with its two effects; edge and shape (Andresen, 2009; A. Fotheringham & Wong, 1991; Maroko et al., 2009; Ord & Getis, 1995; Sadler et al., 2011; Swift et al., 2013; Webster et al., 2006). When subjects of study are sited in larger aggregated geographic units there exists spatial uncertainty and temporal uncertainty, as introduced by Kwan (2012b) as the uncertain geographic context problem (UGCoP). This problem helps researchers know that there are unknowns in the actual areas that exert influences of human behaviour under study and in the time and duration in which individuals are exposed to these neighbourhood influences.

2.2.4 Modifiable Areal Unit Problem

The modifiable areal unit problem (MAUP) coined by Openshaw (1984) refers to the sensitivity of statistical analysis in both the scale used to aggregate spatial observations and the zoning system imposed to contain these observations. It is a pervasive problem when analyzing the relationships between the built environment and health. In this

section, the modifiable area unit problem and the implications of the problem when analyzing the relationship between the built environment and health will be examined. It is argued that both the scale and zonal aspects of MAUP are always 'in-force' when analyzing the relationship between aggregated built environment variables and health.

Rogerson (2001) reports that Gehlke and Biehl (1934) found that when analyzing census data, the correlation coefficients increase with increasing levels of geographic aggregation. Therefore, larger numbers of small sized census areas reveal smaller correlation coefficients than small numbers of large census areas. As the size of the aggregations increase it is possible that interesting local variations in relationships are 'averaged away' and become unobservable. Openshaw (1984) says that the results of any aggregation of points into areal units might be a function of the size, shape and orientation of the areal units, and thus have more of an influence on the results than the distribution of the points themselves. It is understood that the smaller and more compact the areal unit, the less of a risk of accidentally imposing the modifiable areal unit problem on one's analysis. Many spatial data sets are made up of zones, and the configuration of the zones can affect the outcome of the statistical and interpretive analysis. Figure 2.1, shows two different zoning patterns aggregating the same set of observations. The arrows show migration paths of some populations north to south. In example A, there is a

Figure 2.1 Same migration routes with two zoning systems (Rogerson 2001)
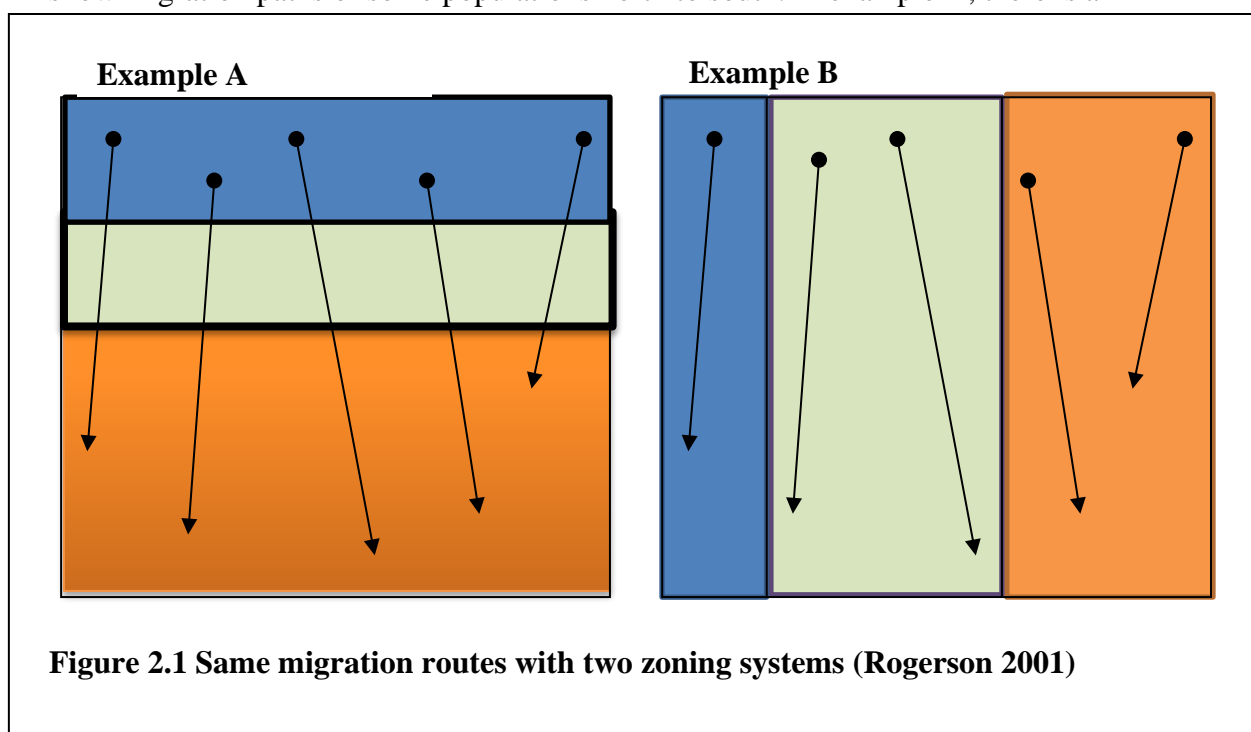
southward migration by crossing several zone boundaries. Example B shows that there was no migration out of the zones, though the migration pattern are the same.

Much of the research of environmental exposure and the influence of the built environment on health is lacking in that much of the previous research often merely counted the number of health-promoting or demoting opportunities within a distance of the individual's perceived location. These opportunity counts were often then used to generate a density metric of accessibility or exposure. These researchers would identify a dependent variable measuring some health status of the individuals which, in turn, was used to generate a statistic identifying some association between the health status of the population and the surrounding environment. Some BE and health researchers have taken on the challenge to minimize and quantify the MAUP in health-related studies (Andresen, 2009; Clark & Scott, 2014; Grady & Enander, 2009; Rainham et al., 2008; Spielman & Yoo, 2009; Swift et al., 2013), while most research only mention that MAUP might be a concern and do not test for its influence.

When the modifiable areal unit is 'in-force' so too is an ecological fallacy (bias). Ecological fallacy is the correlation between individual variables (e.g. health status) generated from a set of deduced variables collected for the group, in this case, the neighbourhood built environment variables to which the individual belongs. The correlation might be false in that it was generated from a larger aggregation but then corresponded to the individual. Ecological fallacy, like the MAUP, is introduced when spatial data is aggregated. It causes significant variation in correlation statistics between exposures and health-related outcomes (Swift et al., 2013). A modifiable areal unit problem sensitivity analysis can be used to investigate the impact of spatial aggregation on the ecological fallacy. Malczewski (1999), suggests a way to disaggregate spatial data into a rectangular grid-based set of isotropic tessellations and experiment by then re-aggregate these tessellations to test for the effect of the MAUP. It is with this effort that researchers can identify the aggregations that keep the interesting local variations from being averaged out. Some researchers, particularly in environmental sciences and ecology have been experimenting with the use of hexagonal tessellations as a way to model, monitor, and sample across the earth's surface at multiple scales (Birch et al., 2007; J.

Gilliland & Olson, 2013; J. Gilliland et al., 2011; Gregory et al., 2008; Sahr, 2008; Zhou et al., 2013)

The implications are clear for researchers investigating the relationships between the built environment and health. Firstly, researchers should not assume arbitrary boundaries and large aggregations of spatial phenomena will not impose MAUP and by consequence the ecological fallacy. Secondly, a sensitivity analysis should be performed by disaggregating and re-aggregating isotropic spatial units to test for bias. Thirdly, researchers should consider an approach that maps the movement of an individual across the landscape so that exposure can match more closely to that individual.

2.2.5 Uncertain Geographic Context Problem

In the ecological approach to health research, it is well understood that environmental exposure has an association to health effects, but the associations can be multi-causal and probabilistic (Krieger, 1994; Lalonde, 1974; Ozonoff, 1994). The ecological model is best suited to help take into account the complex social and spatial contexts (structure) within which every individual exists and how individual behaviour (agency) is influenced by these structures, and in return how individuals can exert influence on these structures through individual choice, education and policy interventions (Egger & Swinburn, 1997).

It is common, in spatial epidemiological research, that contextual spatial units (neighbourhoods) are used as the method to assign area-based attributes to populations to examine the effects of the area-based attributes on individual health behaviours and health outcomes (Brownson et al., 2009). Each spatial context is an aggregation, in some way, of the attributes of the geographic features representing health promoting or health damaging opportunity structures.

In these studies (Apparicio et al., 2008; J. Gilliland et al., 2006; J. Gilliland & Ross, 2005; J. A. Gilliland et al., 2012; Macintyre et al., 2002; Miles et al., 2008; J. Pearce et al., 2006; Thierry et al., 2013; Tucker et al., 2009), opportunity structures are defined as those places in, and the socio-economic factors of, a neighbourhood that are theorized to be associated to the individual health outcomes of the sample population within that

neighbourhood. By setting boundaries or spatial extent limits to both the opportunity structures and to the sample population the modifiable areal unit problem (MAUP) is imposed with both zonal and scalar effects (Openshaw, 1984), while also imposing the boundary problem with its shape and edge effects (Andresen, 2009; A. Fotheringham & Wong, 1991; Maroko et al., 2009; Ord & Getis, 1995; Sadler et al., 2011; Swift et al., 2013; Webster et al., 2006). Therefore, all of the various methods of neighbourhood delineation impose uncertainty and ecological fallacy (Kwan, 2012b). These additional problems must be identified and repaired in what was coined by Kwan (2012b) as the uncertain geographic context problem (UGCoP). The methodological issues that the UGCoP could impose on spatial epidemiological research are vital and could lead to inferential errors about the associations observed.

2.2.6 Mitigating UGCoP with GIS and GPS

The challenge of using GIS alone to meet the spatial complexity and temporal issues arising from uncertain geographic context problem is daunting. Recent research, has combined the use GIS and GPS technologies to address like problems (Duncan et al., 2013; Elgethun et al., 2007; Han et al., 2013; Jones et al., 2009; Kim et al., 2012; Loebach & Gilliland, 2014; Mavoa et al., 2011; M. Oliver et al., 2010; Rainham et al., 2008). A wearable GPS device tracks and records where and when individuals travel through space and time. In this way, the path and temporal nature of the GPS tracks are known. The spatial resolution provided by the GPS is the best way to generate the correct geographic context of a subject thus saving researchers the futility of trying to conceptualize boundaries of the real spatial context. However, the coordinate point clouds generated by the GPS can be difficult to interpret. As a response to the complexity of the mass of data generated by GPS, studies have been conducted to generate area-based features from the GPS points by using standard deviational ellipses as a way to generate boundaries (Boruff et al., 2012; Loebach, 2013; Rainham et al., 2010).

Studies have employed GIS analysis techniques to associate the structures in the spatial context to the *individual-level* GPS tracks. The combination of GIS spatial layers and the GPS tracks have opened the door for further research in associating exposure to health-promoting structures and the duration of exposure (engagement) at those structures.

There are some limits to this approach in that there is no direct information about how the individual is using these structures and if exposure is occurring at all.

The information gaps when using GIS and GPS could be filled using a mixed method approach of both quantitative and qualitative analyses to understand the social interactions of the subjects (Kwan, 2012a, 2012b). Loebach and Gilliland (2014) included diaries in combination with individual neighbourhood mapping exercises as a way to help bridge the gap.

When a GPS based research design or participatory study is not realistic due to time or resource constraints then areal interpolation (Cai et al., 2006; Flowerdew & Green, 1992; Goodchild et al., 1993; Haining, 2009; Henry & Boscoe, 2008; Malczewski, 1999; Ratcliffe, 2004; Reibel, 2007; Rushton et al., 2006; Swift et al., 2013) and dasymetric mapping (Holt et al., 2004; Mennis, 2003) techniques have been employed to reduce the problem with the spatial structural complexity of the problem. Areal interpolation techniques and dasymetric mapping techniques can be used to reshape, resize, and re-proportion variables within spatial boundaries. The area-based variables would be either be removed, reorganized, or re-proportioned into their corresponding smaller spatial context areas, in a way disaggregating the spatial context to make a better model of reality. It is expected with this approach that the MAUP and UGCoP (but not in the temporal context) could be mitigated in some way.

2.3   Conclusion

The literature review was divided into two parts. The primary purpose of part one was to show the existing evidence that demonstrates that there is an association between children's health and environmental influencers, with a specific focus on the benefits of 'being outdoors'. The stage was then set to identify the relevancy of the two methodological studies within the dissertation. The second part focused on the issue of 'uncertainty' in geographic analyses, with particular consideration of the implications of uncertainty in spatial data, data classification and spatial analyses when mapping human subjects and the built environment. We saw that it is common practice in recent GIS studies of accessibility and exposure that home location (i.e., home address) proxies (e.g.,

census tracts, dissemination areas, postal codes) used to represent a subject's location. It was also shown that the use of large areal units to represent neighbourhoods is common. In the first study, presented in Chapter 3 of this thesis, we will see how choosing an inappropriate address proxy and large areal units can bias spatial measurements and skew both distance and classification statistics.

2.4   References

Aalders, H. J. (2002). The registration of quality in a GIS. In W. Shi, P. F. Fisher, & M. Goodchild (Eds.), *Spatial Data Quality* (pp. 186-199). London: Taylor & Francis.

Albrecht, J. (2007). *Key Concepts and Techniques in GIS*. New York: Sage.

Amoly, E., Dadvand, P., Forns, J., Lopez-Vicente, M., Basagana, X., Julvez, J., . . . Sunyer, J. (2014). Green and blue spaces and behavioral development in Barcelona schoolchildren: the BREATHE project. *Environ Health Perspect, 122*(12), 1351-1358. doi:10.1289/ehp.1408215

Andresen, M. A. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography, 29*(3), 333-345. doi:10.1016/j.apgeog.2008.12.004

Anselin, L. (2006). How (not) to lie with spatial statistics. *Am J Prev Med, 30*(2 Suppl), S3-6. doi:10.1016/j.amepre.2005.09.015

Apparicio, P., Abdelmajid, M., Riva, M., & Shearmur, R. (2008). Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. *Int J Health Geogr, 7*, 7. doi:10.1186/1476-072X-7-7

Ball, K., Timperio, A., Salmon, J., Giles-Corti, B., Roberts, R., & Crawford, D. (2007). Personal, social and environmental determinants of educational inequalities in walking: a multilevel study. *J Epidemiol Community Health, 61*(2), 108-114. doi:10.1136/jech.2006.048520

Birch, C., Oom, S., & Beecham, J. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling, 206*(3-4), 347-359. doi:10.1016/j.ecolmodel.2007.03.041

Bjork, J., Albin, M., Grahn, P., Jacobsson, H., Ardo, J., Wadbro, J., . . . Skarback, E. (2008). Recreational values of the natural environment in relation to neighbourhood satisfaction, physical activity, obesity and wellbeing. *Journal of Epidemiology & Community Health, 62*(4), e2-e2. doi:10.1136/jech.2007.062414

Boruff, B., Nathan, A., & Nijënstein, S. (2012). Using GPS technology to (re)-examine operational definitions of 'neighbourhood' in place-based health research IJHG 11-22.pdf>. *Int J Health Geogr, 11*(12).

Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: state of the science. *Am J Prev Med, 36*(4 Suppl), S99-123 e112. doi:10.1016/j.amepre.2009.01.005

Brusilovskiy, E., Klein, L. A., & Salzer, M. S. (2016). Using global positioning systems to study health-related mobility and participation. *Soc Sci Med, 161*, 134-142. doi:10.1016/j.socscimed.2016.06.001

Burgi, R., Tomatis, L., Murer, K., & de Bruin, E. D. (2016). Spatial physical activity patterns among primary school children living in neighbourhoods of varying socioeconomic status: a cross-sectional study using accelerometry and Global Positioning System. *BMC Public Health, 16*, 282. doi:10.1186/s12889-016-2954-8

Buttenfield, B. (1991). A rule for describing line feature geometry. In B. Buttenfield & R. McMaster (Eds.), *Map generalization: Making rules for knowledge representation* (pp. 150-239). New York: Wiley & Sons.

Cai, Q., Bhaduri, B., Coleman, P., Rushton, G., & Bright, E. (2006). Estimating small-area populations by age and sex using spatial interpolation and statistical inference methods. *Transactions in GIS, 10*, 577-598.

Carlson, C., Aytur, S., Gardner, K., & Rogers, S. (2012). Complexity in built environment, health, and destination walking: a neighborhood-scale analysis. *J Urban Health, 89*(2), 270-284. doi:10.1007/s11524-011-9652-8

Carlson, J., Jankowska, M., Meseck, K., Godbole, S., Natarajan, L., Raab, F., . . . Kerr, J. (2015). Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. *Med Sci Sports Exerc, 47*(3), 662-667. doi:10.1249/MSS.0000000000000446

Chuang, Y. C., Cubbin, C., Ahn, D., & Winkleby, M. A. (2005). Effects of neighbourhood socioeconomic status and convenience store concentration on individual level smoking. *J Epidemiol Community Health, 59*(7), 568-573. doi:10.1136/jech.2004.029041

Clark, A., & Scott, D. (2014). Understanding the Impact of the Modifiable Areal Unit Problem on the Relationship between Active Travel and the Built Environment. *Urban Studies, 51*(2), 284-299. doi:10.1177/0042098013489742

Cleland, V., Crawford, D., Baur, L., Hume, C., Timperio, A. F., Salmon, J., . . . Salmon, J. (2008). A Prospective Examination of Children's Time Spent Outdoors, Objectively Measured Physical Activity and Overweight. *International Journal of Obesity, 32*, 1685-1693. doi:10.1038/ijo.2008.171

Cooper, A. R., Page, A. S., Wheeler, B. W., Hillsdon, M., Griew, P., & Jago, R. (2010). Patterns of GPS measured time outdoors after school and objective physical activity in English children: the PEACH project. *Int J Behav Nutr Phys Act, 7*, 31. doi:10.1186/1479-5868-7-31

Couclelis, H. (2003). The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS, 7*(2), 165-175. doi:10.1111/1467-9671.00138

David, B., Van Den Herrewegen, M., & Salgé, F. (1996). Conceptual models for geometry & quality of geographic information. In P. A. P. A. Burrough & U. Frank (Eds.), *Geographic Objects with Indeterminate Boundaries* (pp. 193-206). London: Taylor & Francis.

Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., & Duncan, M. J. (2013). Portable global positioning system receivers: static validity and environmental conditions. *Am J Prev Med, 44*(2), e19-29. doi:10.1016/j.amepre.2012.10.013

Egger, G., & Swinburn, B. (1997). *An "ecological" approach to the obesity pandemic* (Vol. 315).

Elgethun, K., Fenske, R. A., Yost, M. G., & Palcisko, G. (2002). Time-Location Analysis for Exposure Assessment Studies of Children Using a Novel Global Positioning System Instrument. *Environ Health Perspect, 111*(1), 115-122. doi:10.1289/ehp.5350

Elgethun, K., Yost, M. G., Fitzpatrick, C. T., Nyerges, T. L., & Fenske, R. A. (2007). Comparison of global positioning system (GPS) tracking and parent-report diaries to characterize children's time-location patterns. *J Expo Sci Environ Epidemiol, 17*(2), 196-206. doi:10.1038/sj.jes.7500496

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Front Public Health, 2*, 36. doi:10.3389/fpubh.2014.00036

Faber Taylor, A., & Kuo, F. E. M. (2011). Could Exposure to Everyday Green Spaces Help Treat ADHD? Evidence from Children's Play Settings. *Applied Psychology: Health and Well-Being, 3*(3), 281-303. doi:10.1111/j.1758-0854.2011.01052.x

Flouri, E., Midouhas, E., & Joshi, H. (2014). The role of urban neighbourhood green space in children's emotional and behavioural resilience. *Journal of Environmental Psychology, 40*, 179-186. doi:10.1016/j.jenvp.2014.06.007

Flowerdew, R., & Green, M. (1992). Developments in Areal Interpolation Methods and GIS. *The Annals of Regional Science, 26*, 67-78.

Fotheringham, A., & Wong, D. (1991). The-modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*, 1025-1044.

Fotheringham, S. A. (1989). Scale-independent spatial analysis. In G. M. & G. S. (Eds.), *Scale-independent spatial analysis*. London: Taylor & Francis.

Frank, L. D., Andresen, M. A., & Schmid, T. L. (2004). Obesity relationships with community design, physical activity, and time spent in cars. *Am J Prev Med, 27*(2), 87-96. doi:10.1016/j.amepre.2004.04.011

Gell, N. M., Rosenberg, D. E., Carlson, J., Kerr, J., & Belza, B. (2015). Built environment attributes related to GPS measured active trips in mid-life and older adults with mobility disabilities. *Disabil Health J, 8*(2), 290-295. doi:10.1016/j.dhjo.2014.12.002

Gilboa, S. M., Mendola, P., Olshan, A. F., Harness, C., Loomis, D., Langlois, P. H., . . . Herring, A. H. (2006). Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environ Res, 101*(2), 256-262. doi:10.1016/j.envres.2006.01.004

Gilliland, J., Holmes, M., Irwin, J., & Tucker, P. (2006). Environmental equity is child's play: Mapping recreational opportunities in urban neighbourhoods. *Vulnerable Children and Youth Studies, 1*(2), 1-13.

Gilliland, J., & Olson, S. (2013). Residential Segregation in the Industrializing City: A Closer Look. *Urban Geography, 31*(1), 29-58. doi:10.2747/0272-3638.31.1.29

Gilliland, J., Olson, S., & Gauvreau, D. (2011). Did Segregation Increase as the City Expanded?: The Case of Montreal, 1881-1901. *Social Science History, 35*(4), 465-503. doi:10.1215/01455532-1381823

Gilliland, J., & Ross, N. (2005). Opportunities for video lottery gambling: An environmental analysis. *Canadian Journal of Public Health, 96*(1), 55-59.

Gilliland, J. A., Rangel, C. Y., Healy, M. A., Tucker, P., Loebach, J. E., Hess, P. M., . . . Wilk, P. (2012). Linking childhood obesity to the built environment: a multi-level analysis of home and school neighbourhood factors associated with body mass index. *Can J Public Health, 103*(9 Suppl 3), eS15-21.

Goldberg, D. W. (2008). *A Geocoding Best Practices Guide*. Springfield, IL: North American Association of Central Cancer Registries.

Goldberg, D. W., Wilson, J. P., Knoblock, C. A., Ritz, B., & Cockburn, M. G. (2008). An effective and efficient approach for manually improving geocoded data. *Int J Health Geogr, 7*, 60. doi:10.1186/1476-072X-7-60

Goodchild, M., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*.

Goodchild, M., & Hunter, G. (1997). A simple positional accuracy measure of linear features. *International Journal Geographical Information Science, 11*(3), 299-306.

Goovaerts, P. (2009). Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. *Spat Spatiotemporal Epidemiol, 1*(1), 61-71. doi:10.1016/j.sste.2009.07.001

Grady, S. C., & Enander, H. (2009). Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology. *Int J Health Geogr, 8*, 10. doi:10.1186/1476-072X-8-10

Gray, C., Gibbons, R., Larouche, R., Sandseter, E. B., Bienenstock, A., Brussoni, M., . . . Tremblay, M. S. (2015). What Is the Relationship between Outdoor Time and Physical Activity, Sedentary Behaviour, and Physical Fitness in Children? A Systematic Review. *Int J Environ Res Public Health, 12*(6), 6455-6474. doi:10.3390/ijerph120606455

Gregory, M. J., Kimerling, A. J., White, D., & Sahr, K. (2008). A comparison of intercell metrics on discrete global grid systems. *Computers, Environment and Urban Systems, 32*(3), 188-203. doi:10.1016/j.compenvurbsys.2007.11.003

Haining, R. (2009). The Special Nature of Spatial Data. In A. Fotheringham & P. Rogerson (Eds.), *Spatial Analysis (Handbook)*. Thousand Oaks, CA: Sage.

Han, D., Lee, K., Kim, J., Bennett, D. H., Cassady, D., & Hertz-Picciotto, I. (2013). Development of Time-location Weighted Spatial Measures Using Global Positioning System Data. *Environ Health Toxicol, 28*, e2013005. doi:10.5620/eht.2013.28.e2013005

Hay, G., Kypri, K., Whigham, P., & Langley, J. (2009). Potential biases due to geocoding error in spatial analyses of official data. *Health Place, 15*(2), 562-567. doi:10.1016/j.healthplace.2008.09.002

Henry, K. A., & Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *Int J Health Geogr, 7*, 3. doi:10.1186/1476-072X-7-3

Holt, J. B., P., L. C., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science, 31*.

Hong, S., & Vonderohe, A. P. (2014). Uncertainty and sensitivity assessments of GPS and GIS integrated applications for transportation. *Sensors (Basel), 14*(2), 2683-2702. doi:10.3390/s140202683

Jacquez, G. M., & Rommel, R. (2009). Local indicators of geocoding accuracy (LIGA): theory and application. *Int J Health Geogr, 8*, 60. doi:10.1186/1476-072X-8-60

Jones, A. P., Coombes, E. G., Griffin, S. J., & van Sluijs, E. M. (2009). Environmental supportiveness for physical activity in English schoolchildren: a study using Global Positioning Systems. *Int J Behav Nutr Phys Act, 6*, 42. doi:10.1186/1479-5868-6-42

Kelly, C. M., Schootman, M., Baker, E. A., Barnidge, E. K., & Lemes, A. (2007). The association of sidewalk walkability and physical disorder with area-level race and poverty. *J Epidemiol Community Health, 61*(11), 978-983. doi:10.1136/jech.2006.054775

Kerr, J., Duncan, S., & Schipperjin, J. (2011). Using global positioning systems in health research: a practical approach to data collection and processing. *Am J Prev Med, 41*. doi:10.1016/j.amepre.2011.07.017

Kerr, J., Marshall, S., Godbole, S., Neukam, S., Crist, K., Wasilenko, K., . . . Buchner, D. (2012). The Relationship between Outdoor Activity and Health in Older Adults Using GPS. *Int J Environ Res Public Health, 9*(12), 4615-4625. doi:10.3390/ijerph9124615

Kerr, J., Norman, G., Godbole, S., Raab, F., & Demchak, B. (2012). Validating GPS data with the PALMS system to detect different active transportation modes. *Med Sci Sport Exer, 44*((5 Suppl)), S2529. doi:10.1249/MSS.0b013e318236a3d2

Kestens, Y., Thierry, B., & Chaix, B. (2016). Re-creating daily mobility histories for health research from raw GPS tracks: Validation of a kernel-based algorithm using real-life data. *Health Place, 40*, 29-33. doi:10.1016/j.healthplace.2016.04.004

Kim, T., Lee, K., Yang, W., & Yu, S. D. (2012). A new analytical method for the classification of time-location data obtained from the global positioning system (GPS). *J Environ Monit, 14*(8), 2270-2274. doi:10.1039/c2em30190c

Klinker, C. D., Schipperijn, J., Christian, H., Kerr, J., Ersbøll, A. K., & Troelsen, J. (2014). Using accelerometers and global positioning system devices to assess gender and age differences in children's school, transport, leisure and home based physical activity. *International Journal of Behavioral Nutrition and Physical Activity, 11*(8).

Klinker, C. D., Schipperijn, J., Kerr, J., Ersbøll, A. K., & Troelsen, J. (2014). Context-specific outdoor time and physical activity among school-children across gender and age: using accelerometers and GPS to advance methods. *Front Public Health, 2*. doi:10.3389/fpubh.2014.00020

Klinker, C. D., Schipperijn, J., Toftager, M., Kerr, J., & Troelsen, J. (2015). When cities move children: development of a new methodology to assess context-specific physical activity behaviour among children and adolescents using accelerometers and GPS. *Health Place, 31*, 90-99. doi:10.1016/j.healthplace.2014.11.006

Krieger, N. (1994). Epidemiology and the web of causation: has anyone seen the spider? *Social Science and Medicine, 39*(7), 887-903.

Kwan, M. (2012a). How GIS can help address the uncertain geographic context problem in social science research. *Annals of GIS, 18*(4), 245-255. doi:10.1080/19475683.2012.727867

Kwan, M. (2012b). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers, 102*(5).

Lalonde, M. (1974). *A New Perspective on the Health of Canadians.* Retrieved from Ottawa, ON:

Lam, M., Godbole, S., Chen, J., Oliver, M., Badland, H., Marshall, S. J., . . . Kerr, J. (2013). *Measuring Time Spent Outdoors Using a Wearable Camera and GPS.* Paper presented at the SenseCam '13 Proceedings of the 4th International SenseCam & Pervasive Imaging Conference, New York.

Larsen, K., & Gilliland, J. (2008). Mapping the evolution of 'food deserts' in a Canadian city: supermarket accessibility in London, Ontario, 1961-2005. *Int J Health Geogr, 7*, 16. doi:10.1186/1476-072X-7-16

Lee, I. M., Ewing, R., & Sesso, H. D. (2009). The built environment and physical activity levels: the Harvard Alumni Health Study. *Am J Prev Med, 37*(4), 293-298. doi:10.1016/j.amepre.2009.06.007

Leslie, E., Saelens, B., Frank, L., Owen, N., Bauman, A., Coffee, N., & Hugo, G. (2005). Residents' perceptions of walkability attributes in objectively different neighbourhoods: a pilot study. *Health Place, 11*(3), 227-236. doi:10.1016/j.healthplace.2004.05.005

Leung, Y., & Yan, J. (1998). A locational error model for spatial features. *International Journal of Geographical Information Science, 12*, 607–620.

Li, F., Fisher, K. J., Brownson, R. C., & Bosworth, M. (2005). Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. *J Epidemiol Community Health, 59*(7), 558-564. doi:10.1136/jech.2004.028399

Loebach, J. (2013). *Children's Neighbourhood Geographies: Examining Children's Perception and Use of Their Neighbourhood Environments for Healthy Activity.* (PhD), University of Western Ontario, London, Ontario.

Loebach, J., & Gilliland, J. A. (2014). Free Range Kids? Using GPS-Derived Activity Spaces to Examine Children's Neighborhood Activity and Mobility. *Environment and Behavior, 48*(3), 421-453. doi:10.1177/0013916514543177

Lovasi, G. S., Weiss, J. C., Hoskins, R., Whitsel, E. A., Rice, K., Erickson, C. F., & Psaty, B. M. (2007). Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree? *Int J Health Geogr, 6*, 12. doi:10.1186/1476-072X-6-12

Macintyre, S., Ellaway, A., & Cummins, S. (2002). Place effects on Health: how can we conceptualise, operation and measure them? *Social Science and Medicine, 55*(1), 125-139.

Maddison, R., Jiang, Y., Vander Hoorn, S., Exeter, D., Ni Mhurchu, C., & Dore, E. (2010). Describing Patterns of Physical Activity in Adolescents Using Global Positioning Systems and Accelerometry. *Pediatric Exercise Science, 22*, 392-407.

Malczewski, J. (1999). *GIS and Multicriteria Decision Analysis*: Wiley & Sons.

Maroko, A. R., Maantay, J. A., Sohler, N. L., Grady, K. L., & Arno, P. S. (2009). The complexities of measuring access to parks and physical activity sites in New York City: a quantitative and qualitative approach. *Int J Health Geogr, 8*, 34. doi:10.1186/1476-072X-8-34

Mavoa, S., Oliver, M., Witten, K., & Badland, H. M. (2011). Linking GPS and travel diary data using sequence alignment in a study of children's independent mobility. *Int J Health Geogr, 10*, 64. doi:10.1186/1476-072X-10-64

Mazumdar, S., Rushton, G., Smith, B. J., Zimmerman, D. L., & Donham, K. J. (2008). Geocoding accuracy and the recovery of relationships between environmental exposures and health. *Int J Health Geogr, 7*, 13. doi:10.1186/1476-072X-7-13

McCracken, D. S., Allen, D. A., & Gow, A. J. (2016). Associations between urban greenspace and health-related quality of life in children. *Prev Med Rep, 3*, 211-221. doi:10.1016/j.pmedr.2016.01.013

Meijles, E. W., de Bakker, M., Groote, P. D., & Barske, R. (2014). Analysing hiker movement patterns using GPS data: Implications for park management. *Computers, Environment and Urban Systems, 47*, 44-57. doi:10.1016/j.compenvurbsys.2013.07.005

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*, 31-42.

Miles, R., Panton, L. B., Jang, M., & Haymes, E. M. (2008). Residential context, walking and obesity: two African-American neighborhoods compared. *Health Place, 14*(2), 275-286. doi:10.1016/j.healthplace.2007.07.002

Oliver, L. N., Schuurman, N., & Hall, A. W. (2007). Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *Int J Health Geogr, 6*, 41. doi:10.1186/1476-072X-6-41

Oliver, M., Badland, H., Mavoa, S., Duncan, M. J., & Duncan, S. (2010). Combining GPS, GIS, and accelerometry: methodological issues in the assessment of location and intensity of travel behaviors. *J Phys Act Health, 7*(1), 102-108.

Openshaw, S. (1984). *The modifiable areal unit problem.* Norwich: Geobooks.

Ord, J., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographic Analysis, 27*(4).

Ozonoff, D. (1994). Conceptions and misconceptions about human health impact analysis. *Environmental Impact Assessment Review, 14*, 499-515.

PALMS. UCSD PALMS Project.   Retrieved from ucsd-palms-project.wikispaces.com

Panter, J. R., Jones, A. P., van Sluijs, E. M., & Griffin, S. J. (2010). Attitudes, social support and environmental perceptions as predictors of active commuting behaviour in school children. *J Epidemiol Community Health, 64*(1), 41-48. doi:10.1136/jech.2009.086918

Patrick, K., Kerr, J., Norman, G., Ryan, S., Sallis, J., & Krueger. (2008). Geospatial measurement & analysis of physical activity: Physical Activity Location Measurement System (PALMS). *Epidemiology, 19*. doi:10.1097/EDE.0b013e318159074b

Pearce, A., P., T., G., & A., C. (2014). Who children spend time with after school: associations with objectively recorded indoor and outdoor physical activity. *Int J Behav Nutr Phys Act, 11*(1).

Pearce, Saunders, D., Allison, P., & Turner, A. (2018). Indoor and Outdoor Context-Specific Contributions to Early Adolescent Moderate to Vigorous Physical Activity as Measured by Combined Diary, Accelerometer, and GPS. *Journal of Physical Activity and Health, 15*, 40-45. doi:10.1123/jpah.2016-0638

Pearce, J., Witten, K., & Bartie, P. (2006). Neighbourhoods and health: a GIS approach to measuring community resource accessibility. *J Epidemiol Community Health, 60*(5), 389-395. doi:10.1136/jech.2005.043281

Personal Activity Location Measurement System User Guide. (2011). UC San Diego.

Quigg, R., Gray, A., Reeder, A. I., Holt, A., & Walters, D. L. (2010). Using accelerometers and GPS units to identify the proportion of daily physical activity located in parks with playgrounds in New Zealand children. *Prev Med, 50*. doi:10.1016/j.ypmed.2010.02.002

Rainham, D., Bates, C. J., Blanchard, C. M., Dummer, T. J., Kirk, S. F., & Shearer, C. L. (2012). Spatial classification of youth physical activity patterns. *Am J Prev Med, 42*(5), e87-96. doi:10.1016/j.amepre.2012.02.011

Rainham, D., Krewski, D., McDowell, I., Sawada, M., & Liekens, B. (2008). Development of a wearable global positioning system for place and health research. *Int J Health Geogr, 7*, 59. doi:10.1186/1476-072X-7-59

Rainham, D., McDowell, I., Krewski, D., & Sawada, M. (2010). Conceptualizing the healthscape: contributions of time geography, location technologies and spatial ecology to place and health research. *Soc Sci Med, 70*(5), 668-676. doi:10.1016/j.socscimed.2009.10.035

Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science, 18*(1), 61-72. doi:10.1080/13658810310001596076

Raustorp, A., Pagels, P., Boldemann, C., Cosco, N., Söderström, F., & Måtensson, F. (2012). Accelerometer measured level of physical activity indoors and outdoors during preschool time in Sweden and the United States. *J Phys Act Health, 9*.

Reibel, M. (2007). Geographic Information Systems and Spatial Data Processing in Demography: a Review. *Population Research and Policy Review, 26*(5-6), 601-618. doi:10.1007/s11113-007-9046-5

Robitaille, E., & Herjean, P. (2008). An analysis of the accessibility of video lottery terminals: the case of Montreal. *Int J Health Geogr, 7*, 2. doi:10.1186/1476-072X-7-2

Rogerson, P. (2001). *Statistical Methods for Geography*. London: Sage.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (2006). Geocoding in cancer research: a review. *Am J Prev Med, 30*(2 Suppl), S16-24. doi:10.1016/j.amepre.2005.09.011

Rutt, C. D., & Coleman, K. J. (2005). Examining the relationships among built environment, physical activity, and body mass index in El Paso, TX. *Prev Med, 40*(6), 831-841. doi:10.1016/j.ypmed.2004.09.035

Sadler, R., Gilliland, J., & Arku, G. (2011). An application of the edge effect in measuring accessibility to multiple food retailer types in Southwestern Ontario, Canada. *Int J Health Geogr, 10*(34).

Sahr, K. (2008). Location coding on icosahedral aperture 3 hexagon discrete global grids. *Computers, Environment and Urban Systems, 32*(3), 174-187. doi:10.1016/j.compenvurbsys.2007.11.005

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Ann Epidemiol, 17*(6), 464-470. doi:10.1016/j.annepidem.2006.10.015

Shi, W., & Liu, W. (2000). A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science, 14*, 51–66.

Spielman, S. E., & Yoo, E. H. (2009). The spatial dimensions of neighborhood effects. *Soc Sci Med, 68*(6), 1098-1105. doi:10.1016/j.socscimed.2008.12.048

Strickland, M. J., Siffel, C., Gardner, B. R., Berzen, A. K., & Correa, A. (2007). Quantifying geocode location error using GIS methods. *Environ Health, 6*, 10. doi:10.1186/1476-069X-6-10

Swift, A., Liu, L., & Uber, J. (2013). MAUP sensitivity analysis of ecological bias in health studies. *GeoJournal, 79*(2), 137-153. doi:10.1007/s10708-013-9504-z

Tandon, P. S., Saelens, B. E., & Christakis, D. A. (2015). Active play opportunities at child care. *Pediatrics, 135*. doi:10.1542/peds.2014-2750

Tandon, P. S., Saelens, B. E., Zhou, C., Kerr, J., & Christakis, D. A. (2013). Indoor versus outdoor time in preschoolers at child care. *Am J Prev Med, 44*. doi:10.1016/j.amepre.2012.09.052

Thierry, B., Chaix, B., & Kestens, Y. (2013). Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *Int J Health Geogr, 12*(14).

Thrift, N. (1985). Of flies and germs: A geography of knowledge. In D. Gregory & J. Urry (Eds.), *Social Relations and Spatial Structure* (pp. 366–403). London: Macmillan.

Tillmann, S., Tobin, D., Avison, W., & Gilliland, J. (2018). Mental health benefits of interactions with nature in children and teenagers: a systematic review. *J Epidemiol Community Health*. doi:10.1136/jech-2018-210436

Tremblay, M. S., Gray, C., Babcock, S., Barnes, J., Bradstreet, C. C., Carr, D., . . . Brussoni, M. (2015). Position Statement on Active Outdoor Play. *Int J Environ Res Public Health, 12*(6), 6475-6505. doi:10.3390/ijerph120606475

Tucker, P., Irwin, J. D., Gilliland, J., He, M., Larsen, K., & Hess, P. (2009). Environmental influences on physical activity levels in youth. *Health Place, 15*(1), 357-363. doi:10.1016/j.healthplace.2008.07.001

Tviete, H., & Langaas, S. (1999). An Accuracy assessment method for Geo line datasets based on buffering. *International Journal of Geographic Information Science, 13*(1), 27-47.

Wan, N., & Lin, G. (2013). Life-space characterization from cellular telephone collected GPS data. *Computers, Environment and Urban Systems, 39*, 63-70. doi:10.1016/j.compenvurbsys.2013.01.003

Wang, J., He, X. G., & Xu, X. (2018). The measurement of time spent outdoors in child myopia research: a systematic review. *Int J Ophthalmol, 11*(6), 1045-1052. doi:10.18240/ijo.2018.06.24

Wang, M. C., Kim, S., Gonzalez, A. A., MacLeod, K. E., & Winkleby, M. A. (2007). Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. *J Epidemiol Community Health, 61*(6), 491-498. doi:10.1136/jech.2006.051680

Webster, T., Vieira, V., Weinberg, J., & Aschengrau, A. (2006). Method for mapping population-based case-control studies: an application using generalized additive models. *Int J Health Geogr, 5*, 26. doi:10.1186/1476-072X-5-26

Wheeler, B. W., Cooper, A. R., Page, A. S., & Jago, R. (2010). Greenspace and children's physical activity: a GPS/GIS analysis of the PEACH project. *Prev Med, 51*(2), 148-152. doi:10.1016/j.ypmed.2010.06.001

Whitsel, E. A., Quibrera, P. M., Smith, R. L., Catellier, D. J., Liao, D., Henley, A. C., & Heiss, G. (2006). Accuracy of commercial geocoding: assessment and implications. *Epidemiol Perspect Innov, 3*, 8. doi:10.1186/1742-5573-3-8

Wu, J., Jiang, C., Houston, D., Baker, D., & Delfino, R. (2011). Automated time activity classification based on global positioning system (GPS) tracking data. *Environmental Health, 10*.

Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health, 7*, 37. doi:10.1186/1471-2458-7-37

Zandbergen, P. A., & Green, J. W. (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environ Health Perspect, 115*(9), 1363-1370. doi:10.1289/ehp.9668

Zhan, F. B., Brender, J. D., De Lima, I., Suarez, L., & Langlois, P. H. (2006). Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Ann Epidemiol, 16*(11), 842-849. doi:10.1016/j.annepidem.2006.08.001

Zhang, J., & Goodchild, M. (2002). *Uncertainty in Geographical Information.* London: Taylor & Francis.

Zhang, J., & Kirby, R. P. (2000). A geostatistical approach to modelling positional errors in vector data. *Transactions in GIS, 4*(2), 145-159.

Zhou, M., Chen, J., & Gong, J. (2013). A pole-oriented discrete global grid system: Quaternary quadrangle mesh. *Computers & Geosciences, 61*, 133-143. doi:10.1016/j.cageo.2013.08.012

Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics, 64*(1), 262-270. doi:10.1111/j.1541-0420.2007.00870.x

Zimmerman, D. L., Fang, X., & Mazumdar, S. (2008). Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. *Stat Med, 27*(21), 4254-4266. doi:10.1002/sim.3288

Zimmerman, D. L., & Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *Int J Health Geogr, 9*, 10. doi:10.1186/1476-072X-9-10

Zinszer, K., Jauvin, C., Verma, A., Bedard, L., Allard, R., Schwartzman, K., . . . Buckeridge, D. L. (2010). Residential address errors in public health surveillance data: a description and analysis of the impact on geocoding. *Spat Spatiotemporal Epidemiol, 1*(2-3), 163-168. doi:10.1016/j.sste.2010.03.002

Chapter 3

3 Quantifying the magnitude of environmental accessibility misclassification when using imprecise address proxies in public health research

3.1 Introduction

Recent advances in the analytical capacity of desktop geographic information system (GIS) software, combined with the increasing availability of spatially-referenced health and environmental data in digital format, have created new opportunities for making breakthroughs in spatial epidemiology (Zandbergen, 2008). As digital mapping is an abstraction of reality, the spatial data used for visualizing and analyzing geographic phenomena will always be inaccurate to some degree. Such inaccuracies can be compounded when spatially aggregated units are used as locational proxies for mapping and analyzing spatial relationships, rather than more precise geographic locations. In environmental and public health research, it is common to use proxies for sample unit locations, such as centroids of postal/zip codes, census tracts, dissemination areas, blocks, or lots; however, it is very uncommon for studies to address, or even mention, the potential problems ensuing from the positional discrepancies associated with using imprecise address proxies. It is the responsibility of the researcher to identify, quantify, interpret, and attempt to reduce any errors associated with using particular spatial data and locational proxies, so that they do not interfere with any conclusions and recommendations to be made from the findings (Anselin, 2006; Fotheringham, 1989).

Researchers in spatial epidemiology have long been concerned about the absolute or relative spatial accuracy of the address points used to map sample populations or phenomena within a GIS (Goldberg, 2008). Numerous researchers have examined the 'positional errors' which occur when the address from a database is located on a digital map, but the point is not located at the true position of the address (Cayo & Talbot, 2003; Jacquez & Rommel, 2009; Schootman et al., 2007; Strickland et al., 2007; Ward et al., 2005; Zandbergen & Green, 2007). In many previous studies, positional errors are

reported as Euclidian distance errors or errors in the X and Y dimension using the root mean square error (RMSE). While much has been said about positional errors, much less has been said about how study results might be affected when researchers use spatially aggregated units (which themselves might be positionally accurate) as address proxies. Very few studies measure and compare the positional discrepancies between address proxies and the exact address they are used to represent (Bow et al., 2004). A major area of investigation in the fields of spatial epidemiology, health geography, and public health attempts to assess the levels of accessibility or 'exposure' of subject populations to elements in their local environments that are believed to be health-promoting or health-damaging, and are related to certain health-related behaviours or outcomes. Accessibility is typically measured in relation to the distance between subject populations and selected environmental features, and is often operationalized as a binary variable (i.e., accessible/inaccessible, exposed/not exposed) or a density variable (i.e., number of sites within, volume of contaminant within) in relation to an areal unit or 'buffer' of a certain threshold distance (radius) around the subject's address. There is much variability, but unfortunately not much debate, regarding the particular threshold distances to be used in accessibility studies; however, most authors do attempt to justify their choice of threshold distances based on human behavior (e.g. 'walking distance') or perhaps some characteristic of contaminant source (e.g. 150 m from roadway). The chosen accessibility thresholds also typically vary by study population (e.g. children vs. adults), setting (e.g. urban vs. rural), and by health-related outcome (e.g. physical activity vs. asthma). In their study of the environmental influences on whether or not a child will walk or bike to school, for example, Larsen and colleagues (2009) justify the choice of a 1600 m neighborhood buffer based on the local school board cut-off distance for providing school bus service (see also Brownson et al., 2009; Müller et al., 2008; Panter et al., 2010; Schlossberg et al., 2006). Studies which have focussed on access to neighborhood resources such as public parks and recreation spaces have utilized a variety of threshold distances, typically between 400 and 1600 m (compare Bjork et al., 2008; Lee et al., 2007; Maroko et al., 2009; Tucker et al., 2009); however, a threshold distance of 500 m is ideal, as it represents a short 5–7 min walk, therefore easily accessible for populations of all ages (see Sarmiento et al., 2010; Tucker et al., 2009; Wolch et al., 2011). The 5–7 min

walk zone, as represented by the 500 m buffer around a home or public school, is also a common distance used in studies exploring the relationship between access to junk food and obesity (see Austin et al., 2005; Gilliland, 2010; Morland & Evenson, 2009). Studies of 'food deserts' (disadvantaged areas with poor access to retailers of healthy and affordable food) and the potential impact of poor access to grocery stores on dietary habits and obesity have tended to focus on longer distances (800 m or greater), and vary according to urban vs. rural setting (see Larsen & Gilliland, 2008; Pearce et al., 2008; Sadler et al., 2011; Sharkey, 2009; Wang et al., 2007). This analysis focuses on the 10–15 min walk zone (1000m) around a grocery store identified in previous studies of food deserts in Canadian cities (Philippe Apparicio et al., 2007; Larsen & Gilliland, 2008). Research on the role that distance plays from home to emergency services at hospitals shows association with increased risk of mortality with much larger threshold distances than standard 'walk zones' (e.g. greater than 5 km) (see Acharya et al., 2011; Cudnik et al., 2010; Jones & Bentham, 1997; Nicholl et al., 2007). Nicholl and colleagues (2007), for example, discovered that a 10 km increase in straight-line distance to a hospital is associated with a 1% increase in mortality. As hospitals tend to be a regional, rather than a neighbourhood facility, the threshold distance of 10 km will be used for this analysis.

The purpose of this study is to quantify the magnitude of the positional discrepancies and accessibility misclassification that result from using several commonly-used address proxies in public health research. Rushton and colleagues (2006) have argued that when short distances between subject population and environmental features are associated with health effects in epidemiologic studies, the geocoding result must have a positional accuracy that is sufficient to resolve whether such effects are indeed present. Positional errors have been shown to vary significantly by setting (Bonner et al., 2003; Cayo & Talbot, 2003; Ward et al., 2005); therefore, errors are quantified by multiple neighbourhood types: urban, suburban, small town, and rural. 'Meaning' is ascribed to these errors for spatial epidemiologic studies by examining errors in distance and accessibility misclassification concerning several health-related features, including hospitals, public recreation facilities, schools, grocery stores, and junk food retailers.

3.2   Methods

3.2.1   Study area and data

The City of London (population 350,200) and Middlesex County (population 69,024) in Southwestern Ontario, Canada are ideal study areas for examining the geocoding errors in accessibility studies as they encompass a mix of urban, suburban, small town, and rural agricultural areas (Statistics Canada, 2011) (see Figure 3.1). The study area is categorized
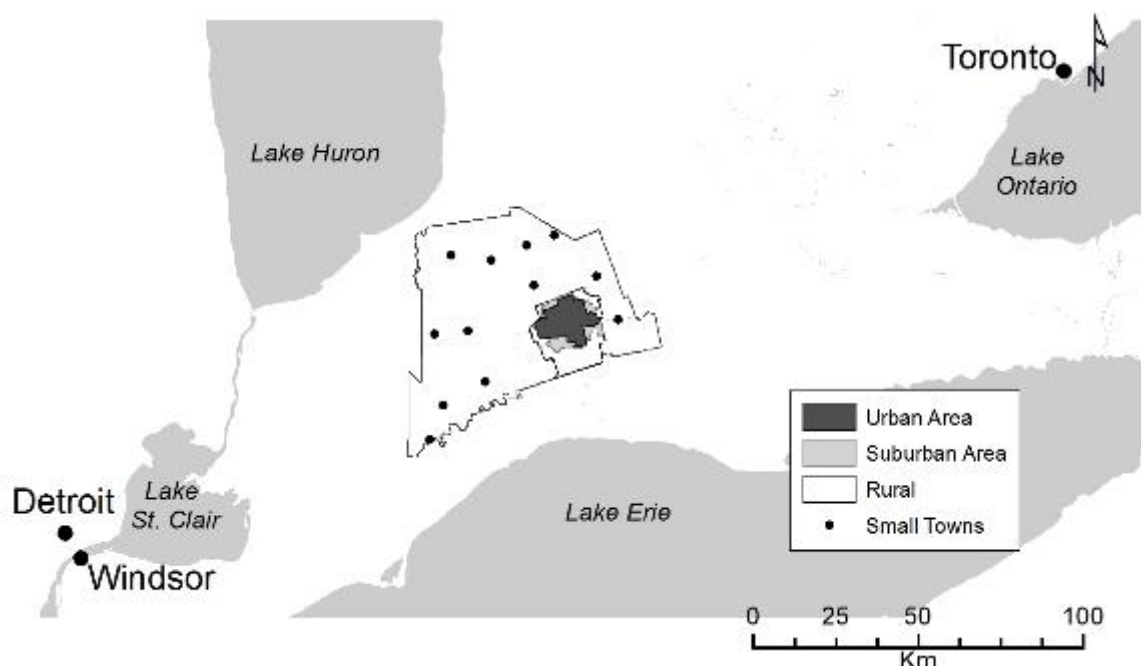
Figure 3.1 Study area: London and Middlesex County, Ontario.

into four neighborhood types as follows: (1) urban areas correspond to neighborhoods in the City of London built primarily before World War II; (2) suburban neighborhoods are areas built following WWII that fall within London's contemporary urban growth boundary; (3) small towns are settlements outside London within Middlesex County, these settlements have fewer than 20,000 inhabitants; and (4) rural areas are defined as all areas of Middlesex County not identified as small town, as well as areas within the city limits of London which are outside its urban growth boundary. All of the areas combine for a total of 104,025 residential addresses, as well as 94 census tracts, 665 dissemination areas and population-weighted dissemination areas, 1410 dissemination blocks, 14,256 postal codes, and 19,365 street segment center points. The spatial relationship between

geographically aggregated units and a sample dwelling centroid illustrated in Figure 3.2.
The dwelling centroid is located within a hierarchical spatial structure starting with the
census tract, moving down to dissemination area, and
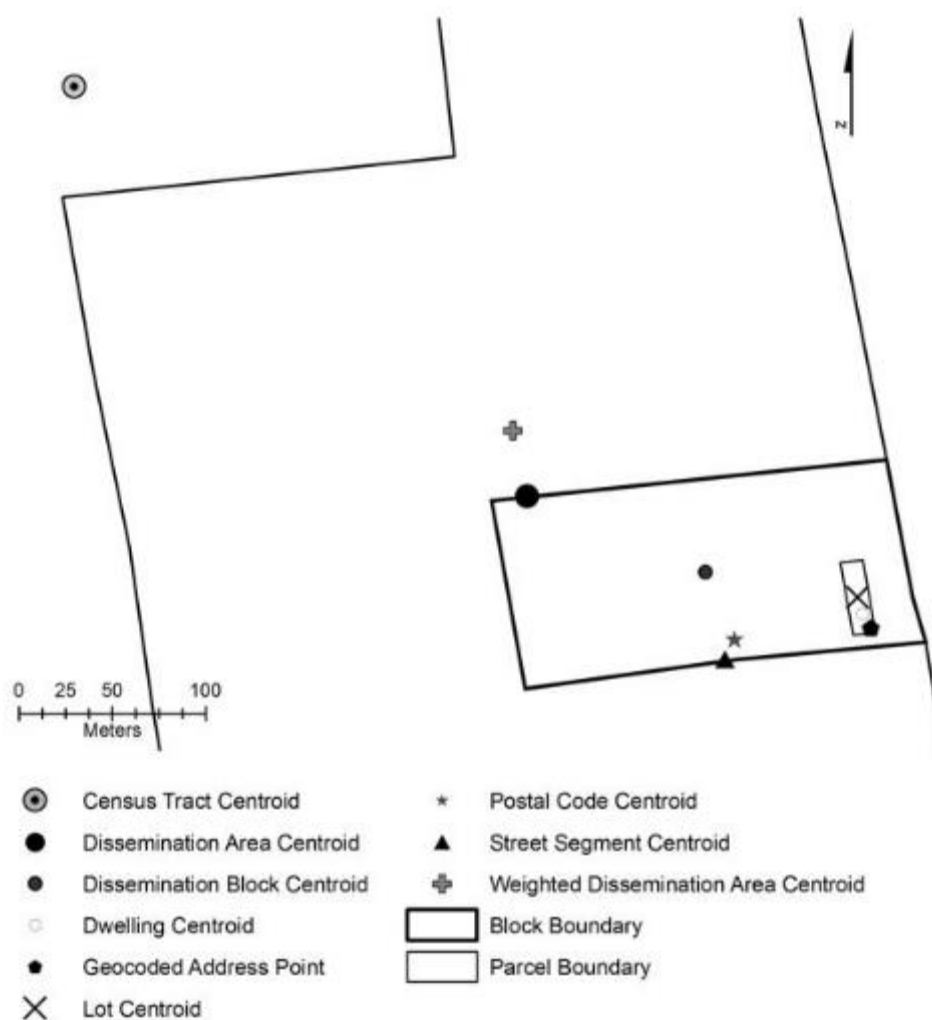
**Figure 3.2 Spatial relationships between various geographic aggregation levels and
their corresponding centroid within a census tract.**

then to the dissemination block and finally the individual parcel of land or lot. The
dwelling unit is also located within a postal code region and on a street segment. Each of
these larger geographic units can be operationalized as point locations according to their
centroids, as seen in Figure 3.2. The hierarchical spatial structure of census data is
organized in such a way that each census tract is made up of multiple contiguous

dissemination areas, which in turn are made up from multiple blocks, which are made from street segments (for the most part, but sometimes divided by rail and natural features). Each street segment is divided as "left" or "right" and is an aggregation of the individual addresses/dwellings on that side of the street.

Digital spatial layers to be used as the address proxies were prepared in ArcMap-ArcInfo10.0 (ESRI, 2011). The census tract, dissemination area, and dissemination block boundary files, supplied by Statistics Canada (2006), were converted to centroids using the 'Feature to Point' tool. These three spatially aggregated units are commonly used in geographic analyses of population data in Canada, and each has tradeoffs for researchers based on the size of the aggregated unit vs. the richness of data available. Dissemination blocks are the smallest of the three geographic units in terms of area; therefore their centroids provide a more spatially accurate proxy for exact address. However, most Canadian census data, except population and dwelling counts, is suppressed at this level, and for this reason, the utility of dissemination blocks in studies of accessibility among population sub-groups is more limited. Dissemination areas are made up of a small group of dissemination blocks. They are commonly-used in population health studies as they are the smallest aggregated geographic unit available for which Statistics Canada releases some key demographic variables (e.g. median household income, population by age, population by ethnicity); nevertheless, a considerable amount of data suppression still occurs at this scale. While census tracts are the most commonly-used proxy for 'neighbourhoods' in sociological, geographical, and population health research in Canada, and they offer the most comprehensive census data for spatial epidemiologic analyses, they are also the largest geographic unit examined in this study. For this reason, they are hypothesized to result in the greatest positional discrepancy when used as address proxies. Additionally, census tracts are only available in metropolitan areas and therefore do not cover most rural areas. The weighted dissemination areas centroids were created using the 'Median Center' tool by leveraging the population distribution data stored within dissemination block centroids which were nested within the dissemination areas. The weighted dissemination areas centroid has been used in previous research (e.g. P. Apparicio et al., 2008; Henry & Boscoe, 2008) and was included in this study as a more representative measure for the probable location of the population within the area. It

is therefore expected to produce a closer approximation for an address proxy than the dissemination area centroid. The postal code boundaries and points were drawn from the Platinum Postal Code Suite (DMTI Spatial Inc., 2009). The typical postal code in a Canadian city is a much smaller geographic unit than the typical US zip code and is commonly used as a proxy for a residential address by Canadian researchers when full civic address is unavailable, or suppressed to maintain subject privacy (e.g. Larsen et al., 2009). The street segment centers were created using the tool 'Feature Vertices to Points' with the CanMap street files (DMTI Spatial Inc., 2009). The geometric center of every street segment was generated as an aggregate address proxy for all the dwellings on that segment. The average street length for rural neighbourhoods was 711 m, 187 m for small towns, 142 m for suburban neighbourhoods, and only 127 m for urban neighbourhoods. All 147,000 addresses points in the study area were supplied by the City and County for every parcel of land, dwelling, business, and institution (City of London, 2010-2013; Middlesex County, 2011). A total of 104,025 address points were identified as residential, and each point was located within the centroid of the dwelling polygons provided by the City and County. A tabular list of each of the residential addresses was generated, and these addresses were used to geocode against the CanMap street files (DMTI Spatial Inc., 2009) using the 'US Address – Dual Ranges' address locator, thus generating interpolated address points with the default 10 m offset from the street centreline. These interpolated addresses, referred to as 'geocoded points' in this paper, are undeniably the most commonly-used address proxies when full address information is available to the researcher. While most researchers use such geocoded points without question, it is argued that even these address proxies could have positional discrepancies which might cause accessibility misclassification and therefore they must also be subjected to further scrutiny. Dwelling centroids are the 'gold standard' of address proxies in this study, to which all other address proxies will be measured. It is the best choice, as all journeys from home begin somewhere within the home building. In this paper, the issues of address validity and match rates for dwelling and lot centroid are controlled for, in that every one of the 104,025 residential addresses were matched at 100%. To calculate accessibility measures, the centroids for dwelling centroids and all the address proxies (except those located on the street segment or a fixed distance from the street segment)

were linked with a connecting lateral line from the proxy address point to the nearest corresponding street segment using a custom algorithm. These lateral lines were included in the network distances reported in the study. The street segment center points already located on the street centerline did not require a lateral line to connect them to the network, while the geocoded points were all standardized to be 10 m from the street centerline and thus the 10 metres were added to the individual distances post process.

GIS layers including the locations of all 6 hospitals, 138 elementary schools, and 512 public recreation spaces within the study area were provided by the geomatics divisions of the City and County (City of London, 2010-2013; Middlesex County, 2011). Addresses for the 52 grocery stores and 1213 junk food retailers (including fast food restaurants and convenience stores) in the study area were provided by the Middlesex-London Health Unit (2010) and geocoded using the master address files provided by the City and County. All data was verified and corrected using orthorectified air photos of London and Middlesex (15 and 30 cm resolution, respectively) (City of London, 2010-2013; Middlesex County, 2011). For built structures, the centroid of the building polygon was used as the address 'gold standard'; however, for recreational places without a defined built structure, such as parks, the access points were manually created using the air photos. The City, County, DMTI Spatial Inc., and Statistics Canada publish no metric regarding the absolute or relative spatial accuracy of their data sets. In this study, the City and County spatial data were accepted as the most spatially accurate of all the data sources. The City and County spatial data were used to create the building centroids for facilities, dwellings, and the centroid for dwelling lots. Spatial features found in the Statistics Canada and DMTI Spatial Inc. data are within 15 m of the same corresponding features in the City and County data for most of the study area. The Statistics Canada and DMTI Spatial Inc. data were used to generate the census tract, dissemination area, weighted dissemination area, dissemination block, postal code centroids, the street segment center, and the geocoded point address proxies, and to generate the shortest path network routes and polygons.

3.2.2 GIS methods

Shortest path routes (by distance) along the street network from the address proxies to the health-related destination facilities were created using the ArcMap 10.0 Network Analyst 'Closest Facility' function (ESRI, 2011). Starting from each dwelling centroid a network route was created to the nearest health-related facility (i.e., the nearest hospital, school, grocery store, junk food outlet, and public recreation facility). This procedure was repeated for every type of health facility until all 104,024 dwelling centroids were assigned a separate shortest path route to one of each of the facility types. The process was then repeated for each of the eight address proxies. The distance measures were stratified into rural, small town, suburban, and urban neighbourhood types and exported from ArcMap 10.0 for analysis in Excel 2010 (Microsoft, 2011) and PASW (SPSS) 18 (IBM Corp., 2011). A recent study of accessibility to multiple food retailer types in rural Middlesex County illustrated how accessibility could be misclassified if facilities outside the county boundary are not considered in distance calculations (Sadler et al., 2011). Sadler and colleagues (2011) demonstrated that when facilities in neighbouring counties were included in the spatial analyses, distance to the nearest grocery store decreased for nearly one-third of households, and distance to nearest fast food outlet decreased for over one-half of households. The edge effect was taken into account in the present study by compiling the datasets for selected health-related facilities in neighbouring counties (within 10 km from the border of Middlesex County) and then including these facilities in the distance calculations.

3.2.3 Misclassified address proxies

When spatial aggregations of the subject populations or geographic features are used as proxies in a study of accessibility, the researcher risks misrepresenting the accessibility metric used in that study. Figure 3.3 illustrates several potential problems of misclassification and miscounting of grocery stores by identifying three accessibility areas; the census tract boundary; a 1000 m network service area buffer originating from the centroid of that same census tract; and a 1000 m network service area buffer originating from a dwelling centroid from within the same census tract. The figure shows that the census tract boundary and the 1000 m network service area buffer around the

census tract centroid does not contain a grocery store, and thus would be coded as inaccessible; however, the dwelling centroid buffer does 'contain' at least one grocery store and would be coded as accessible. Figure 3.3 also illustrates that the count and density metrics will be affected by the positional discrepancy of using imprecise address proxies. We see that the census tract boundary and the buffer around the census tract centroid do not contain any grocery stores, while the dwelling centroid buffer contains two grocery stores. A further look at Figure 3.3 reveals that the distance between the census tract centroid and the dwelling centroid is biased in the direction of the positional discrepancy. In this example, if the census tract centroid were used as the address proxy, the researcher would have coded all sample unit locations within the census tract as not having a grocery store within 1000 m, when in fact, there are two grocery stores within 1000 m for some of the sample units. Moreover, the researcher would have over-estimated the distance to the closest grocery store for many dwelling units, such as the one in the example.
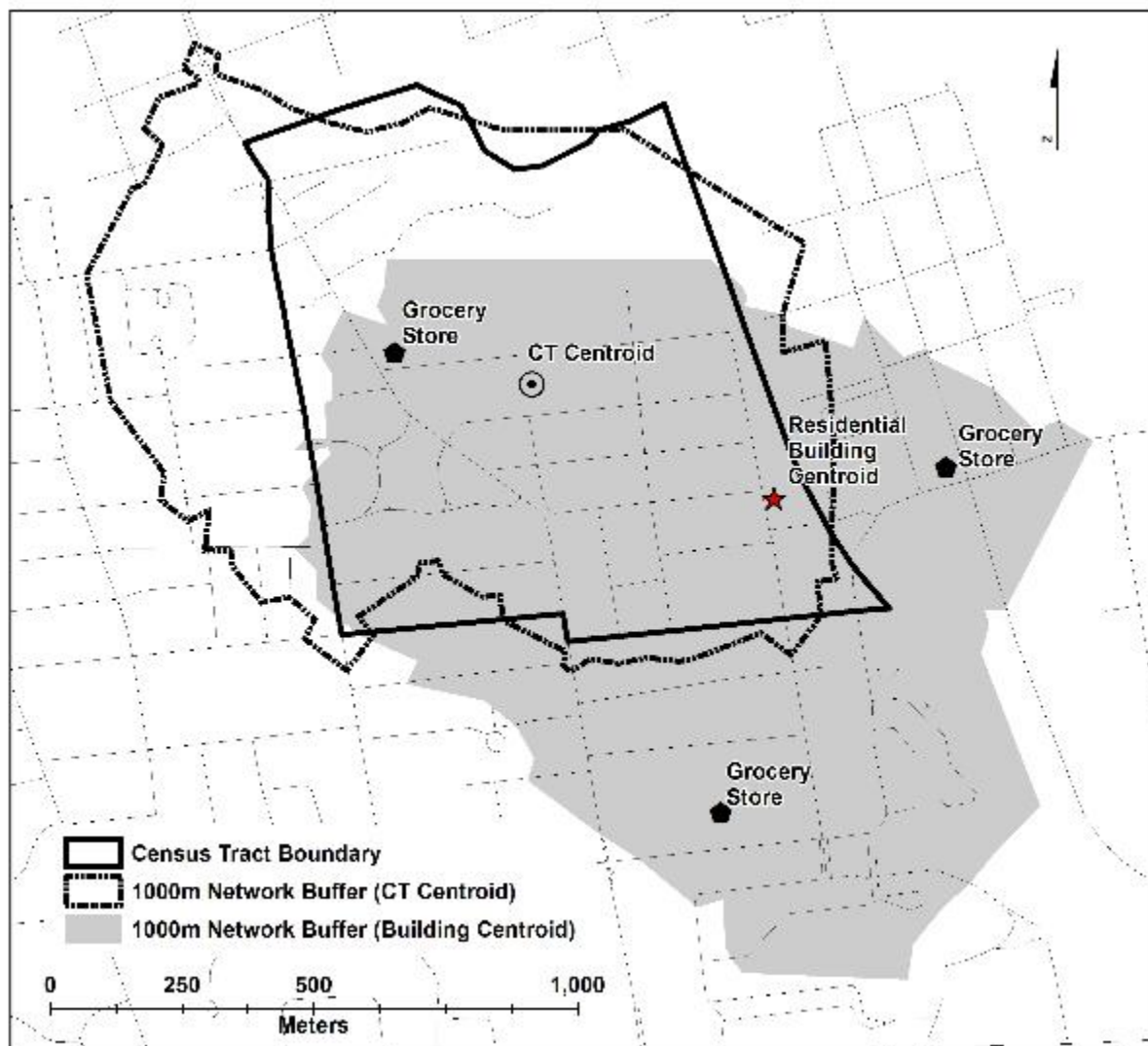
Figure 3.3 Illustration of threshold distance miscoding errors.

Following some commonly-used distances found in previous health-related studies of accessibility (as noted above), the thresholds distances used in this study were: 500 m for junk food and public recreation spaces, 1000 m for grocery stores, 1600 m for schools, and 10 km for hospitals. Shortest path route buffers had been created for each address proxy, and each address proxy point was binary encoded, either the address proxy was inside the threshold (coded as 1) or outside the threshold (coded as 0). The binary variable in matched to every dwelling centroid from every corresponding address proxy and then reported the percentages of improperly coded addresses.

3.2.4   Statistical methods

The distance discrepancies were generated by taking the shortest path distance from a dwelling centroid to a health-related facility and then subtracting the corresponding shortest distance from each corresponding address proxy to that same health facility type.

$$pd = P_{sp} - d_{gs} \tag{3.1}$$

Where the positional discrepancy (*pd*) is the difference between proxy address network shortest path distance (*p_{sp}*) to the closest corresponding health related facility in metres and the "gold standard" centroid of the dwelling shortest path distance to the closest corresponding health related facility (*dgs*).

The Phi correlation coefficient was generated in PASW (SPSS) 18 (IBM Corp., 2011) and was used to measure the association between the binary threshold values (i.e., accessible/inaccessible) between the dwelling centroid threshold value (0,1) to each of its corresponding address proxy threshold values (0,1). Phi will return an association coefficient (Ø) between -1 and +1. A positive value of +1 occurs when all the dwelling threshold values and all the address proxy threshold values are in concordance with one another. Conversely, if there is total discordance between all the dwelling threshold values and all the address proxy thresholds the Phi coefficient will be -1. If some dwelling centroid threshold values differ from those of the corresponding address proxy, the coefficient will begin to move toward 0, thus suggesting a weaker association regarding accessibility encoding for that address proxy. The significantly positive associations (sig. < 0.01) are between 0.7 and 1.0.

3.3   Results

3.3.1   Magnitude of positional discrepancies

In almost every case, urban neighbourhoods show the smallest median distance discrepancies for all address proxies, followed successively by suburban, small town, and rural areas (see Table 3.1). As expected, lot centroids were the most accurate proxy for precise residential dwelling location that were examined in relation to nearest distance to health-related facilities, with the median positional discrepancy (50th percentile) between

lot centroids and dwelling centroids equal to 6–9 m for locations in urban and suburban neighborhoods, 25–43 m for locations in small towns, and 43–50 m for locations in rural

Table 3.1 Median positional discrepancy (metres) by facility type and neighbourhood type.

Neighbourhood type	Rural (m)	Small town (m)	Suburban (m)	Urban (m)
Junk food				
Lot centroids	49	29	9	8
Geocoded point	85	48	51	38
Street segment center	175	65	75	52
Postal code	762	373	78	54
Dissemination block	680	147	127	78
Weighted dissemination area	897	279	168	100
Dissemination area	1054	509	176	113
Census tract	930	1414	243	160
Public recreation places				
Lot centroids	43	43	8	8
Geocoded point	77	34	75	84
Street segment center	185	52	106	102
Postal code	896	1177	114	109
Dissemination block	677	156	176	145
Weighted dissemination area	988	296	228	185
Dissemination area	1070	599	241	207
Census tract	1347	1723	352	247
Grocery stores				
Lot centroids	43	25	6	9
Geocoded point	100	82	80	59
Street segment center	197	95	100	76
Postal code	1196	494	98	79
Dissemination block	810	169	141	112
Weighted dissemination area	1193	335	198	145
Dissemination area	1263	559	201	158
Census tract	1704	1870	373	343
Schools				
Lot centroids	50	32	6	9
Geocoded point	94	51	60	55
Street segment center	173	66	80	66
Postal code	913	711	82	68
Dissemination block	665	148	133	101
Weighted dissemination area	1017	361	187	132
Dissemination area	1140	573	194	140
Census tract	1268	1679	363	251
Hospitals				
Lot centroids	46	27	5	8
Geocoded point	85	65	37	75
Street segment center	187	100	67	93
Postal code	1363	537	78	101
Dissemination block	769	349	176	160
Weighted dissemination area	1350	415	203	166
Dissemination area	1255	538	204	171
Census tract	2088	2166	445	343

areas. The second most accurate proxy for residential location was the geocoded point, with median positional discrepancies between geocoded points and dwelling centroids between 38 and 84 m for residential locations in urban neighborhoods, 37–80 m for locations in suburban neighborhoods, 34–82 m for small-town locations, and 77–100 m

in rural locations. The third most accurate address proxy examined was the street segment centroid, with median positional discrepancies in relation to dwelling centroids between 52 and 102 m for residential locations in urban neighborhoods, 75–106 m for locations in suburban neighborhoods, 52–100 m for small-town locations, and 173–197 metres in rural locations. In urban and suburban areas, the positional discrepancies between postal code centroids and dwelling centroids are very similar to the positional discrepancies between street segment centroids and dwelling centroids; however, the positional discrepancies are drastically worse when using postal codes in small towns (median positional discrepancies between 373 and 1177 m) and rural areas (positional discrepancies between 762 and 1363 m). In rural areas and small towns, the positional errors are always higher when using postal code centroids as address proxies compared to centroids of dissemination blocks, weighted dissemination areas, and dissemination areas. Conversely, postal codes show smaller positional errors than these same address proxies in urban and suburban areas. Census tract centroids are always the address proxy with the most considerable positional error for all neighborhoods and facility types, with median positional discrepancies ranging from the lowest distance error of 160 m (when calculating distance to junk food locations in urban areas) to a high of 2088 m (when calculating distance to hospital in rural areas). Tables 3.2 – 3.6 provide additional information on the positional discrepancies (including mean positional discrepancies, as well as errors at 75th, 90th, 95th, and 99th percentiles) between the address proxies and the dwelling centroids they are meant to represent. The general pattern observable for the median (i.e., 50th percentile) positional discrepancies (reported in Table 3.1) tends to be similar in relative terms, but much less dramatic in terms of absolute positional discrepancies, compared to the mean positional discrepancies, as well as the 75th, 90th, 95th, and 99th percentile of discrepancies.

3.3.2   Positional discrepancy by facility type

The positional discrepancies between the address proxy locations and the dwelling centroids they are to represent not only vary considerably by neighbourhood type but they also vary by health facility type. When lot centroids are used as address proxies, there is a minimal variability between positional discrepancies for all facility types,

regardless of neighbourhood type (rural = ±7 m; urban = ±1 m) (see Table 3.1). Of the 32 unique combinations of address proxies, neighbourhood types, and facility types, it is the junk food outlets (N = 1213) that have the minimum median positional discrepancies 68.8% of the time (22/32), while public recreation facilities (N = 512), singularly, account for almost 50% (15/32) of the facilities with maximum median positional discrepancies. The junk food outlets have small median positional discrepancies for all the address proxies in the urban neighbourhood type. Junk food outlets, also, account for all the minimum median positional discrepancies in suburban and small-town neighbourhood types for postal codes, dissemination block, weighted dissemination area, dissemination area, and census tract proxies. For rural neighbourhoods, the minimum median positional discrepancies for junk food outlets are found when the postal code, weighted dissemination area, and census tract proxies are used. For the most part, public recreation facilities (N = 512) display more considerable median positional discrepancies than all other health-related facilities in urban and suburban areas, while hospitals (N = 6) and grocery stores (N = 52) show the greatest positional discrepancies compared to the other health-related facilities in rural and small towns. The postal code median distance error of 1177 m for a small-town and public recreation facilities is a larger error than rural neighbourhood types and public recreation facilities (896).

Table 3.2 Positional discrepancies (m) from address proxy to closest junk food retailer.

Neighbourhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	69	163	274	1344	984	1325	1415	1427
(n = 16,686)	Median	49	85	175	762	678	897	1054	930
	75th	74	168	370	2040	1431	1930	2033	2159
	90th	166	337	597	3742	2312	3219	3261	3473
	95th	182	471	772	4436	2835	4097	4060	4136
	99th	364	1683	1683	5832	4053	5536	5690	5383
Small town	Mean	38	69	89	1241	455	562	979	1883
(n = 14,139)	Median	29	48	65	373	146	279	509	1414
	75th	35	78	111	1786	458	623	1227	3280
	90th	56	148	181	4467	1231	1207	2528	4190
	95th	99	196	245	5099	2515	2774	3765	4791
	99th	187	351	475	6483	3418	3729	5926	5448
Suburban	Mean	12	83	111	107	186	226	250	297
(n = 54,579)	Median	9	51	75	78	126	168	176	243
	75th	11	76	125	133	255	312	334	423
	90th	17	147	238	224	430	501	564	626
	95th	35	331	380	331	558	637	730	767
	99th	167	551	625	547	881	975	1216	1037
Urban	Mean	13	51	66	71	108	126	139	195
(n = 18620)	Median	8	38	52	54	77	100	113	160
	75th	12	51	81	90	146	176	194	281
	90th	17	77	120	139	230	260	284	405
	95th	30	137	166	187	309	322	351	492
	99th	61	366	377	413	530	527	550	651

n_p – number of address proxies. * Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type

Table 3.3 Positional discrepancies (m) from address proxy to closest public recreation place.

Neighbourhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract[*] (n_p = 94)
Rural	Mean	63	156	270	1645	972	1491	1520	1961
(n = 16,686)	Median	43	77	185	896	677	988	1070	1347
	75th	72	161	401	2393	1427	2177	2180	2749
	90th	158	386	608	4206	2324	3612	3561	4629
	95th	185	606	781	5458	2879	4570	4401	6017
	99th	346	1069	1118	8931	4024	6495	6097	8579
Small town	Mean	41	56	77	1779	503	645	1105	2020
(n = 14,139)	Median	38	34	52	1177	156	296	599	1723
	75th	43	60	92	3109	482	712	1513	3172
	90th	55	109	155	4076	1590	1699	2882	3768
	95th	99	175	235	5095	2770	2971	4010	6521
	99th	195	464	517	9996	3327	4521	6495	7828
Suburban	Mean	11	191	214	211	266	319	347	525
(n = 54,579)	Median	8	75	106	114	176	228	241	352
	75th	9	238	265	257	367	443	473	645
	90th	16	557	586	558	632	732	772	1031
	95th	33	745	766	761	822	920	985	1420
	99th	161	1207	1243	1231	1242	1383	1674	4993
Urban	Mean	11	182	193	195	208	242	265	293
(n = 18,620)	Median	8	84	102	109	145	185	207	247
	75th	12	257	275	279	290	347	377	419
	90th	18	513	527	518	483	523	567	593
	95th	24	632	639	639	608	638	690	714
	99th	60	937	921	953	938	1055	1084	943

n_p – number of address proxies. * Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type

Table 3.4 Positional discrepancies (m) from address proxy to closest grocery store.

Neighbourhood type	% (N = 104024)	Lot (np = 104,024)	Geocoded point (np = 104,024)	Street segment (np = 19,365)	Postal code (np = 14,265)	DB (np = 4210)	Weighted DA (np = 665)	DA (np = 665)	Census tract* (np = 94)
Rural (n = 16,686)	Mean	64	281	377	2000	1095	1707	1721	2581
	Median	43	100	197	1196	810	1193	1263	1704
	75th	74	212	450	2793	1599	2476	2463	3707
	90th	168	568	805	4798	2531	4029	3877	6123
	95th	191	1420	1361	6736	2976	5102	4773	7361
	99th	380	2740	2762	11412	4122	7154	6604	9584
Small town (n = 14139)	Mean	3	115	135	2000	471	651	1102	2730
	Median	25	82	95	494	169	335	559	1870
	75th	31	121	152	3532	493	765	1501	3653
	90th	53	211	288	5529	1465	1623	2963	6821
	95th	95	454	482	8523	2234	2367	3683	9253
	99th	184	567	647	10709	3027	4722	6662	10225
Suburban (n = 54579)	Mean	12	168	197	190	271	327	345	573
	Median	6	80	100	98	141	198	201	373
	75th	9	116	157	162	294	394	404	697
	90th	16	171	258	257	614	727	762	1136
	95th	34	609	736	629	994	1147	1354	1817
	99th	164	2212	2405	2237	2190	2094	2358	3819
Urban (n = 18620)	Mean	11	115	129	132	177	203	217	381
	Median	9	59	76	79	112	145	158	343
	75th	14	88	118	129	209	262	281	553
	90th	19	232	247	274	423	442	476	752
	95th	23	587	594	580	671	656	686	871
	99th	61	854	892	902	924	935	951	1089

np – number of address proxies. * Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type

Table 3.5 Positional discrepancies (m) from address proxy to closest school.

Neighbourhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract[*] (n_p = 94)
Rural	Mean	68	147	254	1547	974	1564	1595	1850
(n = 16,686)	Median	50	94	173	913	665	1017	1140	1268
	75th	76	159	367	2339	1388	2300	2299	2616
	90th	163	294	590	3957	2284	3852	3784	4441
	95th	187	413	743	5021	2929	4795	4752	5550
	99th	378	1074	1071	7693	4060	6308	6303	7493
Small town	Mean	34	65	87	1522	445	666	1087	2155
(n = 14,139)	Median	32	51	66	711	148	361	573	1679
	75th	38	79	108	2465	477	806	1423	2954
	90th	61	115	170	4048	1311	1517	2604	5926
	95th	100	163	228	5922	2271	2723	4322	6875
	99th	189	358	483	6990	3047	4187	6560	7758
Suburban	Mean	13	82	108	109	215	272	300	510
(n = 54,579)	Median	6	60	80	82	133	187	194	363
	75th	10	84	125	136	273	357	379	667
	90th	15	126	191	206	513	609	671	1057
	95th	34	180	286	277	716	838	976	1567
	99th	166	687	713	698	1200	1341	1597	2830
Urban	Mean	13	68	81	84	139	162	171	296
(n = 18,620)	Median	9	55	66	68	101	132	140	251
	75th	14	74	100	110	186	227	241	442
	90th	19	111	147	164	295	331	349	624
	95th	23	170	190	210	387	405	426	724
	99th	61	381	417	409	654	641	651	869

n_p – number of address proxies. * Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type

Table 3.6 Positional discrepancies (m) from address proxy to closest hospital

Neighbourhood type	% (N = 104,024)	Lot (n_p = 104,024)	Geocoded point (n_p = 104,024)	Street segment (n_p = 19,365)	Postal code (n_p = 14,265)	DB (n_p = 4210)	Weighted DA (n_p = 665)	DA (n_p = 665)	Census tract* (n_p = 94)
Rural	Mean	66	176	278	2382	1082	1903	1854	3285
(n = 16,686)	Median	46	85	187	1363	769	1350	1255	2088
	75th	72	284	426	3683	1561	2732	2700	5223
	90th	156	458	655	6150	2508	4496	4400	7815
	95th	180	553	817	8116	3052	5708	5535	9735
	99th	359	859	1148	11812	4375	8419	8292	13483
Small town	Mean	34	178	192	1296	546	674	998	2413
(n = 14,139)	Median	27	65	100	537	349	415	538	2166
	75th	33	335	341	1589	645	832	1273	3266
	90th	56	443	450	3664	1355	1580	2373	5281
	95th	96	511	516	4320	2203	2319	3766	6095
	99th	185	821	828	8095	3060	3690	6689	9435
Suburban	Mean	12	68	93	102	255	287	301	651
(n = 54,579)	Median	5	37	67	78	176	203	204	445
	75th	9	75	127	143	326	384	390	797
	90th	16	178	189	214	556	640	647	1256
	95th	33	188	231	267	777	848	885	1689
	99th	164	367	503	441	1358	1389	1620	5312
Urban	Mean	11	101	104	114	190	207	214	414
(n = 18,620)	Median	8	75	93	101	160	166	171	343
	75th	12	181	170	175	262	292	301	580
	90th	17	193	204	225	380	434	445	835
	95th	22	200	226	263	464	538	555	1078
	99th	58	312	319	362	738	774	814	1668

n_p – number of address proxies. * Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of dwelling centroids; n – number of dwelling centroids by neighborhood type

3.3.3   How positional discrepancy impacts accessibility measures

In addition to reporting the positional discrepancy errors, it is instructive to look at how much of an effect these errors have on the classification of the population aggregated in each of the address proxies. In some health-related accessibility studies continuous variables are used to measure the proximity of health-related facilities to an address proxy. Some studies use binary variables to identify whether or not a health-related facility exists within a set threshold distance (or buffer radius) around a proxy (P. Apparicio et al., 2008; Talen, 2003); still more studies use density and counts, however, as indicated in Figure 3.3, this approach can also lead to severe errors caused by misclassification. Table 3.7 considers the impact of positional discrepancy on accessibility, by reporting the percentage of cases that are incorrectly classified as accessible or not, by address proxy, neighbourhood type and health-related facility type. The general trend is that the smaller the distance threshold, the higher the percentage of

addresses misclassified; also, the larger the geographic area of the unit of aggregation, the higher the percentage of addresses which are misclassified. For example, using the centroid of a large aggregated unit such as a census tract as a proxy instead of a set of residential addresses when calculating containment of a park within 500 m from residential addresses in urban neighbourhoods will result in nearly half (49.5) of all observations being misclassified. On the other hand, using a high threshold distance of 10 km to determine accessibility to hospitals results in no misclassification in urban areas, no matter what the address proxy used (as the threshold practically covers the entire urban area). The Phi coefficient shows a positive association between each of the dwelling centroids and every corresponding address proxy of the coding threshold (inside/ outside) across all the health-related facility thresholds, except for one. There is a weak negative ($\emptyset$ = -0.6, p < 0.01) association for the urban census tract proxy coding thresholds for public recreation facilities. For example, census tract centroids coded as 'outside' (those that do not have a public recreation facility within 500 m) will have many corresponding dwelling centroids coded as 'inside' (those that do have a public recreation facility within 500 m) resulting in this negative association. There is a strong positive association between dwelling centroid and lot centroid for threshold distances of 1 km to grocery stores. If a suburban dwelling centroid is coded as being within 1 km from a grocery store (code = 1), there is a strong probability ($\emptyset$ = 0.996, p < 0.01) that the corresponding lot centroid will also be within 1 km of a grocery store and coded in the same way. Conversely, if a dwelling centroid is coded as being farther away than 1 km from a grocery store (code = 0), then there is the same probability ($\emptyset$ = 0.996, p < 0.01) that the corresponding lot centroid will also be coded in the same way. The range of Phi values for dwellings and corresponding census tracts, dissemination areas, and weighted dissemination area proxies for junk food and recreation places (500 m thresholds) are weakly associated (-0.6 < $\emptyset$ < 0.47, p < 0.01). The fewest misclassification errors and strongest associations for the 500 m thresholds exist for lot centroids ($\emptyset$ > 0.93, p < 0.01) followed by geocoded points (0.6 < $\emptyset$ < 0.87, p < 0.01). Postal code centroids showed very high errors in coding for small town (29.9%) and weak association (rural $\emptyset$ = 0.26, small town $\emptyset$ = 0.29, suburban $\emptyset$ = 0.59, and urban $\emptyset$ = 0.58, p < 0.01).

Table 3.7 Accessibility thresholds: percentage of misclassified observations by address proxy.

Address proxy	Neighbourhood type	Junk food (500 m)	Recreation places (500 m)	Grocery (1 km)	Schools (1.6 km)	Hospitals (10 km)
Census tracts ($N = 94$)	Rural[*] ($n = 17$)	13.5	8.0	4.7	18.3	26.4
	Small town ($n = 3$)	36.7	33.7	21.0	35.7	10.2
	Suburban ($n = 54$)	31.2	47.4	16.8	15.9	5.1[+]
	Urban ($n = 20$)	16.9	49.5	37.1	0.1[+]	0.0[+]
DA(N-665)	Rural ($n = 125$)	7.6	3.7	3.8	11.9	8.6
	Small town ($n = 43$)	35.4	37.1	22.8	29.3	2.7
	Suburban ($n = 367$)	23.9	28.2	11.4	7.6	0.7[+]
	Urban ($n = 130$)	15.5	33.5	15.3[+]	0.1[+]	0.0[+]
Weighted DA ($N = 665$)	Rural ($n = 110$)	9.6	4.7	3.9	11.9	8.5
	Small town ($n = 53$)	31.5	33.5	15.2	19.2	1.2[+]
	Suburban ($n = 372$)	23.0	29.2	11.5	6.7	0.7[+]
	Urban ($n = 130$)	10.7	29.7	15.5[+]	0.1[+]	0.0[+]
DB ($N = 4210$)	Rural ($n = 1499$)	6.9	2.9	2.5	8.5[+]	5.2[+]
	Small town ($n = 593$)	18.2	22.2	11.2[+]	15.3[+]	1.0[+]
	Suburban ($n = 1409$)	18.4	25.6	9.1	5.9[+]	0.9[+]
	Urban ($n = 709$)	12.0	24.5	13.1[+]	1.1[+]	0.0[+]
Postal code ($N = 14,256$)	Rural ($n = 2539$)	9.2	6.8	3.0[+]	8.1[+]	6.9[+]
	Small town ($n = 1003$)	29.9	33.2	27.8	37.0	3.5[+]
	Suburban ($n = 7792$)	11.3[+]	21.0	6.4[+]	2.4[+]	0.3[+]
	Urban ($n = 2922$)	6.5	22.8	10.5[+]	0.1[+]	0.0[+]
Street segment ($N = 19,365$)	Rural ($n = 6310$)	4.3[+]	2.3[+]	1.0[+]	3.6[+]	1.2[+]
	Small town ($n = 2227$)	9.0[+]	8.7[+]	4.3[+]	2.7[+]	0.6[+]
	Suburban ($n = 8364$)	12.4[+]	21.6	6.8[+]	2.3[+]	0.3[+]
	Urban ($n = 2464$)	6.2	23.1	10.1[+]	0.1[+]	0.0[+]
Geocoded ($N = 104,024$)	Rural ($n = 16,686$)	2.9[+]	1.9[+]	1.1[+]	2.5[+]	0.5[+]
	Small town ($n = 14,139$)	7.1[+]	6.7[+]	4.0[+]	2.2[+]	0.4[+]
	Suburban ($n = 54,579$)	8.9[+]	18.3	5.3[+]	1.5[+]	0.2[+]
	Urban ($n = 18,620$)	5.6[+]	21.1	9.4[+]	0.1[+]	0.0[+]
Lot ($N = 104,024$)	Rural ($n = 16,686$)	0.8[+]	0.4[+]	0.2[+]	0.6[+]	0.5[+]
	Small town ($n = 14,139$)	2.0[+]	1.8[+]	0.8[+]	0.6[+]	0.1[+]
	Suburban ($n = 54,579$)	1.7[+]	1.5[+]	0.6[+]	0.4[+]	0.1[+]
	Urban ($n = 18,620$)	1.5[+]	1.7[+]	1.3[+]	0.0[+]	0.0[+]

Abbreviations: DB – dissemination block; DA – dissemination area; N – number of address proxies; n – number of address proxies by neighbourhood type.

[*] Census tracts only exist for rural areas within Census Metropolitan Areas and therefore coverage is biased toward more densely populated rural areas.

[+] Phi coefficient Ø strong positive association (+0.7 to +1.0) sig. < 0.01.

3.4 Discussion

It is common in public health research to use spatially aggregated units as address proxies for the locations of subjects and facilities when more precise address information is unavailable. It is rare, however, for public health researchers to examine, or even mention, the possible distance and misclassification errors resulting from the positional discrepancies between the locations of imprecise address proxies and precise subject locations. It is inappropriate for researchers to ignore these inaccuracies or to merely accept them as an inevitable component of doing spatial research. It is essential to identify and quantify any spatial errors so that we can critically examine research findings and adequately advise those to whom policy recommendations are made regarding the potential correlations between subject populations and accessibility to health promoting/demoting environmental features.

One of the contributions of this study is to quantitatively describe the magnitude of positional discrepancies that result when several of the most commonly-used address proxies are implemented in several different neighbourhood types, including rural, suburban, small town, and urban areas. It is recognized that accessibility thresholds will vary by setting, as well as health outcome or health-related behaviour. Therefore, by demonstrating how the magnitude of the positional discrepancies can affect measures of accessibility (or exposure) to a variety of health-related spaces in different environments and at different distance thresholds, this study also makes a methodological contribution to the environmental and public health literature.

The dwelling as represented by the centroid of the building in which the study participant resides is considered the gold standard for residential address location. If dwelling centroids are not available to the researcher, then the second most accurate address proxy is the centroid of the parcel of land (i.e., the lot) on which the dwelling unit is located; this finding is true regardless of neighbourhood type. When the lot centroid is used as an address proxy, accessibility misclassification errors are virtually nonexistent in urban and suburban neighbourhoods, and are minor in rural areas and small towns.

Where digital files for all residential buildings or residential lots are not available for a study region, but the researcher has access to the complete civic address (i.e., street name and number) for each subject, it is widespread for researchers to geocode their tables of subject addresses using 'address locator' tools to interpolate residential addresses. While the median distance error for this address proxy is too high for researchers to just ignore (ranging from a low of 34 m to a high of 100 m depending on facility and neighborhood types), for the most part, there are few instances of miscoded accessibility when this commonly used address proxy is used: fewer than one-tenth (8.9%) of all observations are misclassified, except for recreation spaces within 500 metres in suburban and urban neighborhoods, where approximately one-fifth of observations are misclassified (18.3% and 21.1%, respectively).

A variation on the interpolated address technique is to use the centroid of the closest street segment as address proxy. This method is useful for environmental equity studies, where researchers may want to map and visualize how access to specific environmental features varies at a fine scale across a study area, but they do not have (or cannot show for privacy reasons) individual address data for subject populations. The street segment centreline address proxy appeared to have fewer distance and misclassification errors than the more commonly-used postal code centroids, particularly for small-town and rural areas.

Postal codes are undoubtedly the most commonly-used proxy for residential addresses of research subjects in Canadian public health studies. In Canada, the postal code centroid is often the best solution when exact addresses are unavailable, or inaccessible due to research ethics board policies and privacy concerns. The results indicate that postal code centroids are reasonably accurate proxies for residential addresses in urban and suburban areas (median positional discrepancies between 54 and 109 m depending on facility type); however, it is recommended that postal codes should be used only with extreme caution for studies based in small town and rural areas of Canada. Positional discrepancies between postal code centroid and dwelling centroid can be very high in rural areas: depending on facility type, median positional discrepancies in rural areas ranged between 762 and 1363 m. Furthermore, postal codes are found to be reasonably

accurate for accessibility studies when distance thresholds are 1000 m or higher; however, it is advised that postal codes should not be used as proxies for residential addresses in accessibility studies where the threshold distances or density buffers are as short as 500 m. Postal code centroids are particularly prone to misrepresenting accessibility in small towns and rural Canada, and therefore should only be used with more caution in spatial epidemiologic research in Canada.

Urban areas show the smallest distance error for all address proxies followed by suburban, small town, and rural neighbourhoods. As expected, the magnitude of positional discrepancies and threshold misclassification errors are more substantial, or most problematic, when the address proxy is the centroid of a large geographic aggregation such as the census tract. In general, the census tract performed poorly as an address proxy except in urban areas where threshold distances are 1600 m or higher. Similarly, it is recommended that centroids of dissemination areas and weighted dissemination areas should only be used as residential address proxies in urban areas when threshold distances are set at greater than 1000 m and in suburban areas when threshold distances are set at greater than 1600 m. As for small Canadian towns, researchers should also avoid all spatially aggregated address proxies for threshold distances less than 1.6 km as the misclassification errors are consistently large, as are the positional discrepancies. While these recommendations are based on the empirical findings related to the specific health-related facilities examined in this study, it is recognized that the positional accuracy required for spatial epidemiology research also depends on the specific acessibility related health outcome under examination (e.g. spatial accuracy is more critical for studies of exposure to air pollution than distance to nearest hospital).

This study examined errors in the shortest path distances from each address proxy to the closest public recreation space, junk food outlet, grocery store, school, and hospital in a full range of neighbourhood types. One way in which this study differs from previous studies of positional error is that street network distances were used in the error calculation, rather than Euclidean distances. Since a subject must use the existing street network (or pathway network) to travel from their dwelling to access the nearest park,

junk food outlet, grocery store, school, or hospital, it would be inaccurate to calculate positional errors and therefore accessibility misclassification as Euclidean or 'crow fly' distances between address proxies and dwelling centroids (except where distances are too small to require use of the network). As a necessary methodological step to create baseline distance measures for comparative purposes, this study assigned health-related accessibility scores to every residential address in the study area. These individual values are at the finest scale so that, in future, they can be aggregated in any geographic frame a researcher would see fit to use. By creating accessibility measures to individual dwelling centroids, researchers are no longer constrained by the (often arbitrary) boundaries of blocks, postal codes, dissemination areas, census tracts, or even counties.

In the last several years, researchers have access to more high resolution address point locations than ever before. In the event that a researcher has a combination of health data, only available at the census tract level, and high resolution address point locations where a distance measure is required, the census tract centroid would not be used. Individual distance measures for individual address point locations within the CT would be calculated, and then those distances would be aggregated to generate an integrated measure for that census tract, in lieu of the CT centroid. All the address proxies identified in this thesis with high positional discrepancies (e.g. census tract, dissemination area, rural postal codes) could have these centroid errors mitigated by combining distances from individual higher resolution address points located within them.

3.5 Conclusion

In spatial epidemiologic and public health research it is common to use spatially aggregated units such as centroids of postal/zip codes, census tracts, dissemination areas, blocks or block groups as proxies for sample unit locations. Few studies, however, have addressed the potential problems associated with using these units as address proxies. Chapter 3 quantifies the magnitude of positional discrepancies and accessibility misclassification that result from using several commonly-used address proxies in public health research. The impact of these positional discrepancies for spatial epidemiology was illustrated by examining the misclassification of accessibility to several health-related facilities, including hospitals, public recreation spaces, schools, grocery stores,

and junk food retailers throughout the City of London and Middlesex County, Ontario, Canada. Positional discrepancies were quantified by multiple neighborhood types, revealing that address proxies are most problematic when used to represent residential locations in small towns and rural areas compared to suburban and urban areas. Findings indicate that the shorter the threshold distance used to measure accessibility between subject population and a health-related facility, the greater the proportion of misclassified addresses. Using address proxies based on large aggregated units such as centroids of census tracts or dissemination areas can result in very large positional discrepancies (median errors up to 343 and 2088 m in urban and rural areas, respectively), and therefore should be avoided in spatial epidemiologic research. Even smaller, commonly-used, proxies for residential address, such as postal code centroids, can have large positional discrepancies (median errors up to 109 and 1363 m in urban and rural areas, respectively) and are prone to misrepresenting accessibility in small towns and rural Canada; therefore, postal codes should only be used with caution in spatial epidemiologic research.

There is a growing trend in public health studies, particularly within the burgeoning field of 'active living research', toward the use of 'ego-centric' units (typically defined by buffers around a study participant's residence) to characterize a participant's neighborhood in order to examine the effect that local environmental factors (e.g. the mix of land uses and coverage of sidewalks) may have on health-related behaviors such as walking (e.g. Larsen et al., 2009) and outcomes such as physical activity levels (Tucker et al., 2009). The findings of this study have revealed that if commonly-used proxies such as centroids of census tracts, dissemination areas, and even postal codes, are used instead of exact addresses, positional discrepancies can be significantly large. If positional discrepancies are large, such 'ego-centric' neighbourhood units will be significantly 'off center', and local environments can be mischaracterized. For example, the chances of misclassifying a health-promoting feature of the neighborhood, such as a park, or a health-damaging feature, such as a junk food outlet, as accessible (or not) can be unacceptably high, particularly when threshold distances are short, such as the commonly-used 500 m buffer (or 5-min walk zone). If positional discrepancies are too large, it will be impossible for the researcher to resolve whether any health effects of an

environment are truly present. Improving the accuracy of the distance calculations increases the utility of the findings so that making decisions and enacting policies aimed at improving a population's spatial accessibility to environmental features will potentially contribute to the overall health and well-being of the population.

3.6   Bridge to Chapter 4

As demonstrated, many researchers have attempted to define accessibility of their subjects to the built environment through the use of address proxies and common geospatial methods. However, despite the wide range of scholarship on the topic, many of these studies suffer from serious methodological weaknesses and inadequate spatial data which, in turn, limit their usefulness. Chapter 3 outlined that the spatial inaccuracies and accessibility misclassifications, when employing some of these common spatial data at the neighbourhood scale, are severe enough to be avoided in future studies. This study was originally published with the title "*Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research*" at a time when those engaged in this type of research were conflating the terms *accessibility* and *exposure*. For the purpose of this thesis the word *exposure* has been replaced with *accessibility* to more closely match the terminology in this thesis and the terminology in the present state of the science.

If researchers are going to suggest and develop appropriate interventions for the design of the built environment, to improve the health of children and the rest of the population, then they must make use of spatially accurate and scale appropriate data. The methods outlined in this chapter apply to other regions in Canada and can be useful for any study documenting geographic accessibility to health-promoting/ health-damaging environmental features. With the foundation of a precise, accurate spatial data and sound geospatial methods, planners and stakeholders can better plan for future interventions.

In Chapter 4, the limits of the ego-centric geospatial methods of distance and network buffers around address proxies are challenged with the incorporation of the GPS tracking of survey participants. By employing GPS, the researcher can measure accessibility as a function of the actual geographic domain rather than arbitrary 'distance from' measures.

The actual places visited and time elapsed at each place presents the opportunity to measure the exposure to and engagement at these places respectively and, therefore, reduces the confounding nature of the modifiable area unit and uncertain geographic context problems, in terms of statistical analysis.

3.7 References

Acharya, A. B., Nyirenda, J. C., Higgs, G. B., Bloomfield, M. S., Cruz-Flores, S., Connor, L. T., . . . Leet, T. L. (2011). Distance From Home to Hospital and Thrombolytic Utilization for Acute Ischemic Stroke. *Journal of Stroke and Cerebrovascular Diseases, 20*(4), 295-301. doi:10.1016/j.jstrokecerebrovasdis.2009.12.009

Anselin, L. (2006). How (not) to lie with spatial statistics. *Am J Prev Med, 30*(2 Suppl), S3-6. doi:10.1016/j.amepre.2005.09.015

Apparicio, P., Abdelmajid, M., Riva, M., & Shearmur, R. (2008). Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. *Int J Health Geogr, 7*, 7. doi:10.1186/1476-072X-7-7

Apparicio, P., Séguin, A.-M., & Naud, D. (2007). The Quality of the Urban Environment Around Public Housing Buildings in Montréal: An Objective Approach Based on GIS and Multivariate Statistical Analysis. *Social Indicators Research, 86*(3), 355-380. doi:10.1007/s11205-007-9185-4

Austin, S. B., Melly, S. J., Sanchez, B. N., Patel, A., Buka, S., & Gortmaker, L. (2005). Clustering of Fast-Food Restaurants Around Schools: A Novel Application of Spatial Statistics to the Study of Food Environments. *American Journal of Public Health, 95*(9), 1575-1581.

Bjork, J., Albin, M., Grahn, P., Jacobsson, H., Ardo, J., Wadbro, J., . . . Skarback, E. (2008). Recreational values of the natural environment in relation to neighbourhood satisfaction, physical activity, obesity and wellbeing. *Journal of Epidemiology & Community Health, 62*(4), e2-e2. doi:10.1136/jech.2007.062414

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Feudenheim, J. L. (2003). Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology, 14*(4), 408-412.

Bow, J., Waters, N., Faris, P. D., Seidel, J. E., Galbraith, D., Knudtson, M. L., . . . Investigators, A. (2004). Accuracy of city postal code coordinates as a proxy for location of residence. *Int J Health Geogr, 3*(5).

Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: state of the science. *Am J Prev Med, 36*(4 Suppl), S99-123 e112. doi:10.1016/j.amepre.2009.01.005

Cayo, M., & Talbot, T. (2003). Positional error in automated geocoding of residential addresses. *Int J Health Geogr, 2*(10).

City of London. (2010-2013). *Parcels, buildings, address points, and health facilities GIS files [DVD]*.

Cudnik, M. T., Schmicker, R. H., Vaillancourt, C., Newgard, C. D., Christenson, J. M., Davis, D. P., & Lowe, R. A. (2010). A geospatial assessment of transport distance and survival to discharge in out of hospital cardiac arrest patients: Implications for resuscitation centers. *Resuscitation, 81*(5), 518-523.

DMTI Spatial Inc. (2009). *Database of postal code centroids and street centerline GIS files [Internet]*.

ESRI. (2011). ArcGIS Desktop (Version 10.0). Redlands, CA: Environmental Systems Research Institute.

Fotheringham, S. A. (1989). Scale-independent spatial analysis. In G. M. & G. S. (Eds.), *Scale-independent spatial analysis.* London: Taylor & Francis.

Gilliland, J. (2010). The Built environment and obesity: trimming waistlines through neighbourhood design. . In Bunting, Filion, & Walker (Eds.), *Canadian cities in transition.* (4th ed., pp. 391–410): Oxford Univ Press.

Goldberg, D. W. (2008). *A Geocoding Best Practices Guide.* Springfield, IL: North American Association of Central Cancer Registries.

Henry, K. A., & Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *Int J Health Geogr, 7*, 3. doi:10.1186/1476-072X-7-3

IBM Corp. (2011). IBM PASW Statistics for Windows (Version 18.0). Armonk, NY: IBM Corp.

Jacquez, G. M., & Rommel, R. (2009). Local indicators of geocoding accuracy (LIGA): theory and application. *Int J Health Geogr, 8*, 60. doi:10.1186/1476-072X-8-60

Jones, A. P., & Bentham, G. (1997). Health service accessibility and deaths from asthma in 401 local authority districts in England and Wales, 1988–92. *Thorax, 52*(3), 218-222.

Larsen, K., & Gilliland, J. (2008). Mapping the evolution of 'food deserts' in a Canadian city: supermarket accessibility in London, Ontario, 1961-2005. *Int J Health Geogr, 7*, 16. doi:10.1186/1476-072X-7-16

Larsen, K., Gilliland, J., Hess, P., Tucker, P., Irwin, J., & He, M. (2009). The influence of the physical environment and sociodemographic characteristics on children's mode of travel to and from school. . *American Journal of Public Health, 99*(3), 520-526. doi:10.2105/AJPH.2008

Lee, R. E., Cubbin, C., & Winkleby, M. (2007). Contribution of neighbourhood socioeconomic status and physical activity resources to physical activity among women. *Journal of Epidemiology Community Health, 61*(10), 882-890.

Maroko, A. R., Maantay, J. A., Sohler, N. L., Grady, K. L., & Arno, P. S. (2009). The complexities of measuring access to parks and physical activity sites in New York City: a quantitative and qualitative approach. *Int J Health Geogr, 8*, 34. doi:10.1186/1476-072X-8-34

Microsoft. (2011). Excel (Version 2010). Redmond, WA: Microsoft Inc.

Middlesex-London Health Unit. (2010). *Database of food retailers [DVD]*.

Middlesex County. (2011). *Database of parcels, address point, aerial photos, and health facilities GIS files [DVD]*.

Morland, K. B., & Evenson, K. R. (2009). Obesity prevalence and the local food environment. *Health Place, 15*(2), 491-495. doi:10.1016/j.healthplace.2008.09.004

Müller, S., Tscharaktschiew, S., & Haase, K. (2008). Travel-to-school mode choice modelling and patterns of school choice in urban areas. *Journal of Transport Geography, 16*(5), 342-357. doi:10.1016/j.jtrangeo.2007.12.004

Nicholl, J., West, J., Goodacre, S., & Turner, J. (2007). The relationship between distance to hospital and patient mortality in emergencies: an observational study. *Emerg Med J, 24*(9), 665-668. doi:10.1136/emj.2007.047654

Panter, J. R., Jones, A. P., van Sluijs, E. M., & Griffin, S. J. (2010). Attitudes, social support and environmental perceptions as predictors of active commuting behaviour in school children. *J Epidemiol Community Health, 64*(1), 41-48. doi:10.1136/jech.2009.086918

Pearce, J., Hiscock, R., Blakely, T., & Witten, K. (2008). The contextual effects of neighbourhood access to supermarkets and convenience stores on individual fruit and vegetable consumption. *J Epidemiol Community Health, 62*(3), 198-201. doi:10.1136/jech.2006.059196

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (2006). Geocoding in cancer research: a review. *Am J Prev Med, 30*(2 Suppl), S16-24. doi:10.1016/j.amepre.2005.09.011

Sadler, R., Gilliland, J., & Arku, G. (2011). An application of the edge effect in measuring accessibility to multiple food retailer types in Southwestern Ontario, Canada. *Int J Health Geogr, 10*(34).

Sarmiento, O. L., Schmid, T. L., Parra, D. C., Díaz-del-Castillo, A., Gómez, L. F., Pratt, M., . . . Duperly, J. (2010). Quality of Life, Physical Activity, and Built Environment Characteristics Among Colombian Adults. *Journal of Physical Activity and Health, 7*(S2), S181-S195.

Schlossberg, M., Greene, J., Paulsen Phillips, P., Johnson, B., & Parker, B. (2006). School Trips: Effects of Urban Form and Distance on Travel Mode. *Journal of the American Planning Association, 72*(3), 337-346.

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Ann Epidemiol, 17*(6), 464-470. doi:10.1016/j.annepidem.2006.10.015

Sharkey, J. R. (2009). Measuring potential access to food stores and food-service places in rural areas in the U.S. *Am J Prev Med, 36*(4 Suppl), S151-155. doi:10.1016/j.amepre.2009.01.004

Statistics Canada. (2006). *Census boundary files [Internet]*.

Statistics Canada. (2011). *Rural and Small Town Canada Analysis Bulletin 2011*. Ottawa.

Strickland, M. J., Siffel, C., Gardner, B. R., Berzen, A. K., & Correa, A. (2007). Quantifying geocode location error using GIS methods. *Environ Health, 6*, 10. doi:10.1186/1476-069X-6-10

Talen, E. (2003). Neighborhoods as Service Providers: A Methodology for Evaluating Pedestrian Access. *Environment and Planning B: Planning and Design, 30*(2), 201-218.

Tucker, P., Irwin, J. D., Gilliland, J., He, M., Larsen, K., & Hess, P. (2009). Environmental influences on physical activity levels in youth. *Health Place, 15*(1), 357-363. doi:10.1016/j.healthplace.2008.07.001

Wang, M. C., Kim, S., Gonzalez, A. A., MacLeod, K. E., & Winkleby, M. A. (2007). Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. *J Epidemiol Community Health, 61*(6), 491-498. doi:10.1136/jech.2006.051680

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., . . . Hartge, P. (2005). Positional Accuracy of Two Methods of Geocoding. *Epidemiology, 16*(4), 542-547.

Wolch, J., Jerrett, M., Reynolds, K., McConnell, R., Chang, R., Dahmann, N., . . . Berhane, K. (2011). Childhood obesity and proximity to urban parks and recreational resources: a longitudinal cohort study. *Health Place, 17*(1), 207-214. doi:10.1016/j.healthplace.2010.10.001

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems, 32*(3), 214-232. doi:10.1016/j.compenvurbsys.2007.11.006

Zandbergen, P. A., & Green, J. W. (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environ Health Perspect, 115*(9), 1363-1370. doi:10.1289/ehp.9668

Chapter 4

4 Objectively measuring children's time spent outdoors exposure to, and engagement in green space while using passive GPS devices

4.1 Introduction

Recently, there has been significant amounts of literature published on the health benefits of spending time outdoors, especially for children. Research shows that spending time outdoors can positively impact children's physical activity, mental health and well-being, and social, and cognitive development (Gilliland, 2018). Dramatic increases in sedentary behaviour and time spent using electronic devices is a concern of many parents, practitioners, and researchers, which supports further investigation into the relationship between time spent outdoors and a variety of children's health outcomes.

A body of the literature identified in a recent systematic review assessed the effect of outdoor time on children's physical activity, sedentary behaviour, and physical fitness (Gray et al., 2015). Several studies included in this review confirmed that there were overall positive effects of outdoor time on physical activity, sedentary behaviour, and cardiorespiratory fitness (Gray et al., 2015). Each of the studies assessing physical activity also found higher rates of physical activity outdoors as compared with rates indoors. Distinguishing the amount of time that children spend outdoors is important since studies have shown that children are more physically active in outdoor environments (Raustorp et al., 2012).

Previous research has conceptualized interactions with nature outdoors into three types: accessibility, exposure, and engagement (Tillmann et al., 2018). In their review of the impacts of nature interactions on children's mental health, they describe how some of these studies using measures of accessibility, such as residential proximity to outdoor greenspace, are limited in their assessment of children's interaction with their natural environments, as accessibility studies do not establish that children are making use of these spaces. Research therefore should be accounting for the individual choices made by

children when discussing interactions with particular environments. Exposure to, and engagement with outdoor environments give us more accurate representations of the actual time children spend at particular locations.

One of the significant barriers to assessing children's time spent in outdoor environments is the inability to precisely determine whether a child's GPS tracks are indoors or outdoors. Previous research has used time blocks, such as a school schedule, to determine whether a child is most likely indoors or outdoors (Loebach & Gilliland, 2014). However, this has limitations in that it assumes that all participants are in similar environments indoors (in the school building) and outdoor (school yard) based on a school schedule and does not take into account trips away from school, or absences. Some studies have supplemented time blocks with self or parent-reported activity diaries detailing where children's activities are taking place; however, this again makes some assumptions based on time blocks included in the diary and creates an opportunity for inaccurate reporting by children due to recall bias (Loebach & Gilliland, 2014). Studies that rely on self-reporting alone or on parents reporting on behalf of the child have also been used to classify use or time spent in specific spaces which again leaves room for inaccurate reporting, as well as not being an accurate representation of every space a child interacts with on a daily basis (Amoly et al., 2014; Faber Taylor & Kuo, 2011; Flouri et al., 2014; McCracken et al., 2016). Being able to accurately determine whether a single GPS point is indoors or outdoors is of crucial importance for understanding children's activities, and ultimately, for understanding the link between environments and health.

Ellis et al. (2014) identified 89.9% accuracy when using the random forest model with 150 hours of GPS tracks when predicting active travel behaviours for two of their adult research assistants after filtering the data through the PALMS tool. Meanwhile, Wu et al. (2011) found the random forest less useful for identifying outdoor travel times when using only the raw GPS tracks using derived acceleration rate, speed, distance difference between subsequent readings, and distance ratio rather than any National Marine Electronic Association (NMEA) quality metrics.

The purpose of the analysis in this chapter is to code, with some certainty, any discrete coordinate computed by a wearable GPS as to whether the subject wearing the GPS unit was indoors or outdoors at the time of coordinate generation.

Presently, there are lack of tools to accurately quantify time spent in indoors or outdoors for children (Wang et al., 2018) and unreliable tools to assign those activities meaning in terms of exposure to and engagement in those spaces. This study is carried out within a socio-ecological framework that recognizes that there are many types of influence on children's behaviours and health outcomes (Sallis et al., 2006; Stokols, 1992). It is the purpose of this study to (1) develop a tool that accurately designates whether an individual GPS point is generated as indoors or outdoors and (2) assess the accuracy of this tool through evaluating a random sample of GPS points, (3) assign environmental and neighbourhood variables to the outdoor activities, and (4) identify statistically significant metrics against a set of socioeconomic indicators to answer: where and for how long do children engage outdoors, and what neighbourhood and socioeconomic environment might be aiding or hindering this behaviour?

4.2   Methods

The methods section is divided into three parts. The first part will describe the study design, instruments, and GPS data collection from the multi-year research study entitled the Spatio-Temporal Environment and Activity Monitoring (STEAM) Project. The second part of the methods section will describe (a) the classification algorithm used in predicting, with certainty, all coordinates generated by a wearable GPS as to whether the subject wearing the GPS was indoors or outdoors at the time of coordinate generation, and (b) the binning method used to combine the outdoor GPS tracks with the built environment variables as a step to reduce both the modifiable areal unit problem (MAUP) (Openshaw, 1984) and the uncertain geographic context problem (Kwan, 2012b) effects. The third part of this section describes the statistical methods used to report on the STEAM participants' exposure to and engagement in different outdoor environments.

4.2.1 Spatio-Temporal Environment and Activity Monitoring (STEAM) Protocol

4.2.1.1 Study design

The Spatio-Temporal Environment and Activity Monitoring (STEAM) protocol has been employed three times since 2010 by the Human Environments Laboratory in the Department of Geography and Western University. The larger STEAM study utilizes a mixed-methods approach to understand how children aged 9-14 years engage with different health-promoting or health-damaging environmental features in their neighbourhoods. The study presented in this chapter utilizes data gathered during the first two phases of the larger study (i.e. STEAM I and STEAM II). The age of the participants in the STEAM project corresponds to critical life stage when children are independently mobile and have a growing sense of their own environments (Rissotto & Tonucci, 2002).

STEAM I and II were designed to examine the potential causal effects of the built environment on children's health-related behaviours in the Southwestern Ontario region of Canada. Before launching both studies, approvals were obtained from the Non-medical Research Ethics Board of the University of Western Ontario (see Appendix A and B); approval was also obtained from the regional school boards to approach schools for participation. Seven elementary schools in the City of London, four urban and three suburban, participated during the first two years of the study (STEAM I), while 30 schools participated in the second study (STEAM II), representing populations from London, and local municipalities. In this Chapter, only STEAM I participants were used to study exposure and engagement.

Children in grades 5 to 8 (approximately 8 to 14 years of age) in each school were considered eligible to participate. All children who received permission to participate from a parent or guardian, and who signed their own Child Assent Form, were allowed to participate in the study.

4.2.1.2 GPS data collection

For STEAM I, participants at seven elementary schools completed a 7-day multi-tool protocol to document their neighbourhood activities, movements and experiences.

Participating children (n=220) wore portable GPS monitors (VGPS-900 by Visiontac) shown in Figure 4.1, during all waking hours for up to 7 days during the winter-spring season (February-March) and again in the spring-summer season (April to June); GPS units marked a spatial coordinate for each second the unit was in use. Participants also completed detailed daily activity and travel diaries, and both children and parents completed comprehensive surveys on children's neighbourhood activities, environmental perceptions and mobility behaviours. In STEAM II, during the study period (2011-2013), participants (n=946) at the thirty elementary schools also completed a 7-day multi-tool protocol similar to STEAM I, but with the two seasons being fall (October-November) followed by spring (March-April). In both studies, the GPS devices were attached to a lanyard and the children were instructed to wear the device around their neck from the time they rose in the morning until bed-time (see Figure 4.2).

Figure 4.1: VGPS-900 GPS receiver with lanyard

Figure 4.2: Example of a child wearing VGPS-900 GPS receiver
(niece of the Author – not a STEAM participant)

The protocol of all the STEAM projects required that researchers enter the schools every day that the children were present (outside of weekends) so that each GPS device could have the data downloaded and checked for malfunctions. Each GPS device was programmed by the researchers to generate a coordinate every second and to include a set of signal quality metrics on each coordinate. The Visiontac VGPS-900 GPS receiver devices were used in both studies. The accuracy and precision of the VGPS-900 GPS receiver was reported by the manufacturer using the Circular Error Probability (CEP) metric. The CEP reports distances from a true position on a horizontal plane (accuracy). It includes a probability statistic (percentage) of the actual GPS measurements falling within a distance to the true position. The middle of the cross-hairs in Figure 4.3 represent the true position. The distance value in the CEP metric is illustrated as the radius of a circle (precision). When the radius is small and centred on the cross-hairs the GPS will report with higher accuracy and higher precision. The reported precision and accuracy of the VGPS-900 GPS receiver during optimal conditions (full unobstructed sky view) and wide augmentation assisted system (WAAS) enabled (Differential GPS - DGPS) is 1.5m CEP (30%-50%) p=0.05, 2.5m CEP (95%) p=0.05, and when WAAS correction not in effect (Non-Differential GPS – Non-DGPS) the expected accuracy is 3m CEP (30%-50%) p=0.05, 5m CEP (30%-50%) p=0.05. In other words, in optimal conditions the VGPS-900 GPS receiver will measure within 1.5 metres of its true position 30 to 50% of the time, and will measure within 2.5 metres of its true position 95% of the time (19 times out of 20). The precision of the GPS measurements shown as circles in Figure 4.3 are a function of 3-D trilateration calculations computed by the receiver from the satellite radio signals downrange that are being received (pseudorange). If the sky view is clear, and there are no blocking structures and there are 4-plus satellites well-spaced above the horizon, then the positional dilution of precision (PDOP) would be small. A small dilution of precision will likely generate a truer measurement. The VGPS-900 was evaluated against other GPS units including the Qstarz brand GPS and was chosen due to the accuracy of the recorded positions during testing and its larger storage capacity so that 1 second epochs could be employed.

Low Accuracy
Low Precision

High Accuracy
Low Precision

Low Accuracy
High Precision

High Accuracy
High Precision
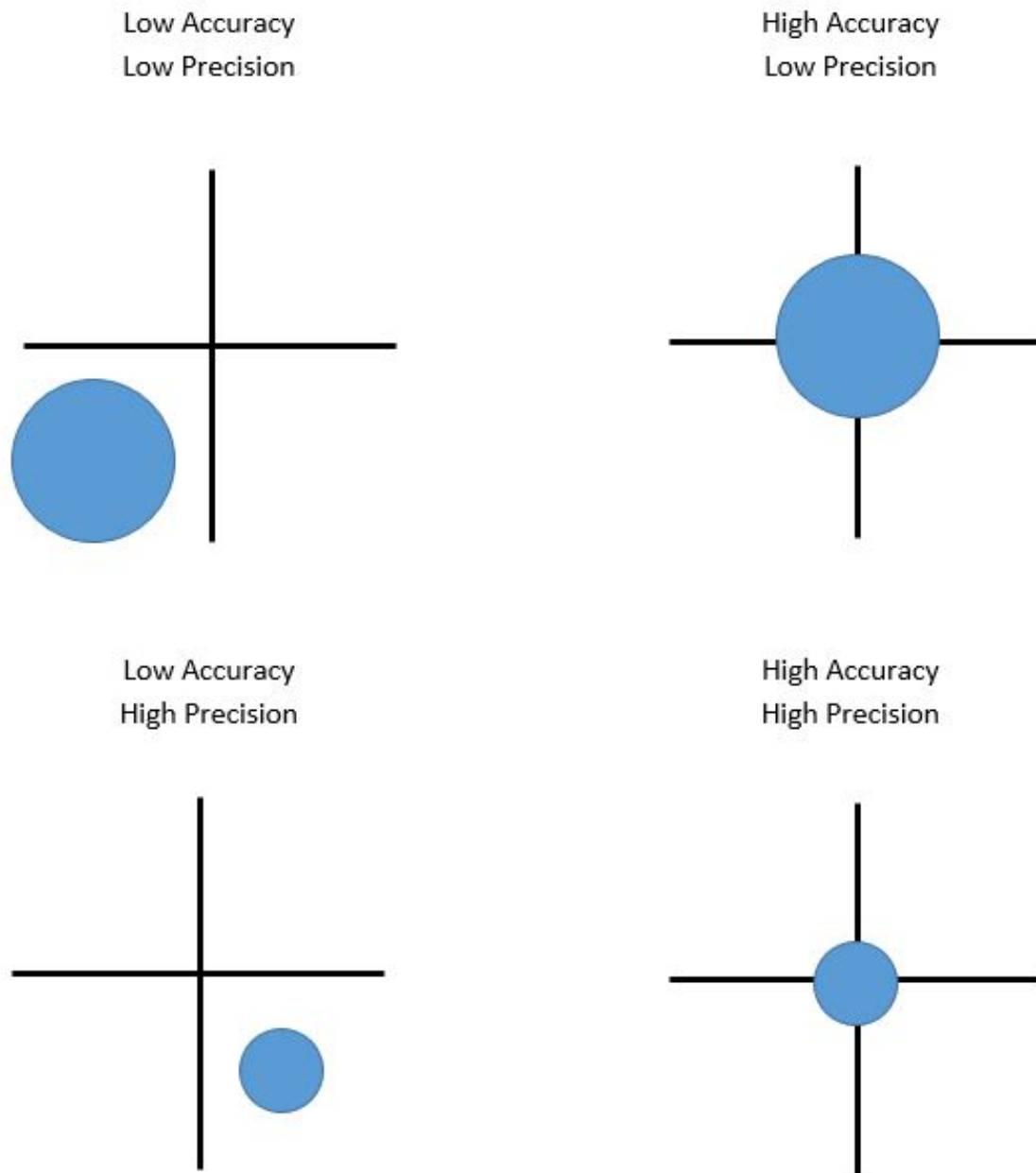
Figure 4.3: GPS accuracy vs precision

The accuracy and precision of the GPS was objectively measured by placing the GPS receiver at a City of London Engineering Survey Monument - Horizontal Control Monument number 028941099 as shown in Figure 4.4.

Figure 4.4: Horizontal Survey Monument

The monument used was bronze cap type encased in concrete and had a known horizontal locational accuracy to the millimeter using the North American Datum 1983 (NAD83) Zone 17 coordinate system and a non-geodetic accuracy of a decimeter in the vertical. The monument is situated on a bridge with a full view of the sky in all directions which is considered the best case for measuring the accuracy of the device. The VGPS-900 GPS receiver was placed on a surveyor's tripod at a height of 1.2 m above the monument as shown in Figure 4.5.

Figure 4.5: Author aligning surveyor tripod over horizontal survey monument

The GPS was configured to measure coordinates at 1 second epochs for 30 minutes (1800 coordinates). For accuracy, the difference between the GPS measured coordinates and Horizontal Control Monument coordinate was calculated in ArcGIS and used to generate the Root Mean Square Error (RMSE) for the GPS device:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \qquad (4.1)$$

where the square root of the sum of the square mean difference between each of the predicted values (*P*) of Horizontal Control Monument location (X and Y) and the observed values (*O*) – Vision Tac GPS device (X and Y) of the set of values (n = 1800).

Precision is expressed as the degree of spread in the GPS points. It is measured by first identifying the mean centre of the 1800 GPS points in ArcGIS Pro 2.x (ESRI, 2018). The *Mean Center* tool creates a single point location where the sum of squared Euclidean distances between that mean centre point and each GPS point in the sample are minimized. The distance from each GPS point to the mean centre point was calculated and used to generate the precision (spread) reported as a standard deviation.

The GPS receiver used in STEAM I and II supports a combination on National Marine Electronic Association (NMEA) version 3.01 (NMEA, 2018) data sentences including the GGA (Global Positioning System Fix Data), GLL (Geographic Position, Latitude / Longitude and time), GSA (GPS DOP and active satellites), and HDT (heading in degrees North). These include, for each timestamp, the PDOP (Positional Dilution of Precision, HDOP (Horizontal Dilution of Precision), VDOP (Vertical Dilution of Precision), Speed, Height, and qualitative accuracy values ('2DGPS' and '3DGPS'). The GPS quality metrics including PDOP, HDOP, and VDOP, along with the Height and Speed were stored during the STEAM study and, as we shall see later in this chapter, were imperative in the classification of indoors/outdoors.

Throughout each study the STEAM researcher team copied daily sets of GPS tracks from each GPS device. At the end of each week in the study, researchers combined the full week of GPS data into a digital spreadsheet by participant and imported the data into ArcGIS 10.x (ESRI, 2018) for visual inspection, data quality metrics, data formatting and cleaning. The data was then exported to the GIS and stored as tables in a relational database management system.

4.2.1.3   Past research using STEAM I and STEAM II

Several other graduate students have used STEAM data to investigate how children's environments influence their health-related behaviours. Topics included healthy eating

(Rangel, 2013), sleep (McIntosh, 2014), active transportation (Hill, 2012; Fitzpatrick, 2013; Richard, 2014; Rivet 2016), neighbourhood mobility and activities (Loebach, 2013), physical activity (Richard, 2014; Mitchell, 2016) and health-related quality of life (Tillmann, 2018). Previous STEAM researchers used GPS tracking and built environment variables to identify children's activity spaces and/or routes between home and school, and then associated them with a health-related behavior such as active travel (e.g. Rivet, 2016). To identify outdoor activity, these previous projects either visually inspected the GPS tracks against aerial photography and vector-based ancillary GIS data, or used a combination of visual inspection matched against activity diaries. This study differs from previous outdoor researchers using STEAM in that the outdoor data filtering and classification uses a rigorous method combining the use of kernel based algorithm identifying routes and stops with a random forest classifier to identify outdoor generated GPS points.

4.2.2   Data filtering and classification

4.2.2.1   SphereLab Tool: Activity place detection algorithm for GPS data

Thierry et al. (2013), developed a tool called the "Activity place detection algorithm for GPS data", hereby called the *SphereLab Tool*. This tool uses a kernel density approach to filter the GPS points to identify places where the participant stopped for some defined duration and to generate routes to and from these locations. The tool was validated on a study by Kestens et al. (2016) to build individual mobility histories with 95.8% of the GPS points correctly classified as an activity location of a trip route. Empirically, it was found that for the STEAM GPS data, a kernel distance of 75 m was optimal for identifying the stops from a cluster of routes. It was found that 100% of the stops generated during school hours landed on school property, with all of them within 20m of the school building.

There were three drawbacks with using the tool. The first was that the code had not been updated since the year 2013 and was only written to work in ArcGIS 10.1(ESRI, 2011). The tool was unsupported and failed to process fully using more recent versions of

ArcGIS. This author was then required to edit the published open source python programming code to upgrade the software to work in more recent versions of ArcGIS 10.4-10.6. The second drawback was that the format requirements for the tool dictated that the archived STEAM I and II GPS data needed to be further processing. Additional coding was necessary to prepare the STEAM GPS points to meet the tool requirements (UTC Time field, WGS84 Projection, and a removal of the most grievously erroneous GPS points). Additionally, the output thematic data (routes and stops) generated by the tool did not match the needs of this project which was remedied through the creation of a custom "wrapper" tool the modified S*phereLab Tool's* programming code in the middle. The custom tool pre-processed the STEAM GPS data for input into the modified S*phereLab Tool* and post-processed the output of the S*phereLab Tool* to suit the needs of this study.

The *SphereLab Tool* creates both route (vector polylines) and stops (points) GIS data. In this study, the custom tool was run for each participant by season, and by day so that daily routes and stops could be recorded for each participant. The routes polylines were post-processed to assign the route with the STEAM participant ID, the duration of the route, the start/end point IDs, and the start and end times on the routes. The thematic data of the stop points were similarly post-processed to include the STEAM participant ID, duration, and stop ID (corresponding to the route start/end point IDs). The key goal for the output was to maintain the link to and from the original GPS points with the routes and stops prior to indoor/outdoor classification steps discussed in Section 4.2.5 of this Chapter.

4.2.2.2   Hexagon tessellated surface

An isotropic hexagonal tessellation of the City of London was built in ArcGIS to provide an avenue for overcoming the potential effects of the modifiable areal unit problem (MAUP) (Openshaw, 1984) and the uncertain geographic context problem (UGCoP) (Kwan, 2012b) and with their corresponding potential ecological fallacies. An array of equal sized hexagons, as seen in Figure 4.6, made of six 10m sides (60m perimeter) and a 20m wide diameter, with an area of 259.8m^2, was created over the entire city (see Figure 4.7). The zonal effect after attribution, with any rotation or re-orientation of the surface

would be slight because the hexagon orientation reverts to its original form when rotated at 60° increments. The hexagon orientation looks the same after every 60° rotation, in comparison to the standard square (e.g. raster cell), which looks the same when rotated at 90° increments. The compactness ratio of a hexagon, as compared to the geometry of a

Diameter: 20m
Perimeter: 80m
Area: 400m^2
Compactness: 0.79

20m

Diameter: 20m
Perimeter: 60m
Area: 259.8m^2
Compactness: 0.91
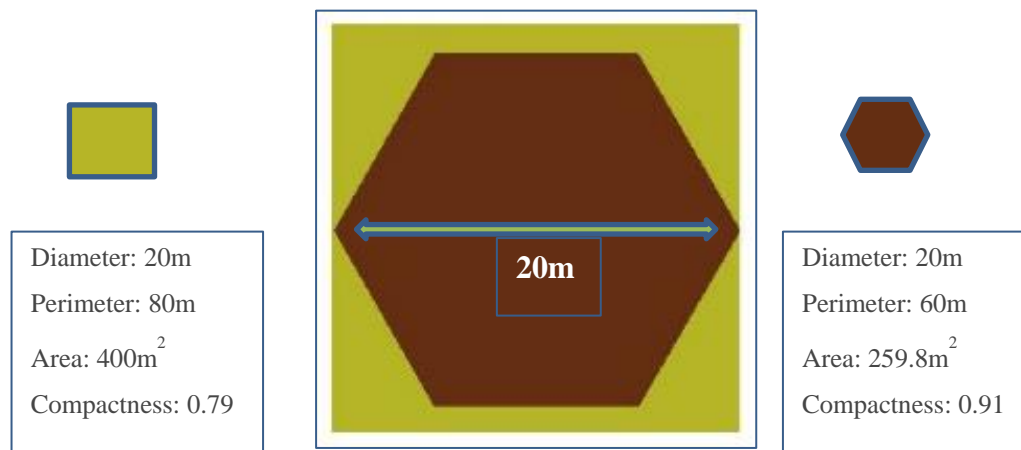
Circle Compactness = 1.0

Figure 4.6: 20m hexagon tessellation

circle, is (0.91) which is a more compact shape than that of a standard square (0.79), and is therefore the most compact shape which can be used for a 2D surface tessellation.

Spence and White (1992), as reported in Davis and Robinson (2012) produced a hexagonal tessellation of the landscape for the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EPA-EMAP) of the United States. They reported that an array of hexagons provides contiguous spatial coverage that is isotropic, and less likely to be coincident with features such as jurisdictional boundaries, buildings or roads (see Figure 4.7).

Figure 4.7: 20m hexagon tessellation

The area-based environmental variables were disaggregated, reorganized, and re-proportioned into their corresponding hexagon context areas to make a better model of reality (see Figures 4.8 and 4.9). The use of 20 metre diameter hexagons to store the important built environment variables related to children's activity outdoors is theorized to be small enough to represent a spatial extent to which a child, who finds themselves located inside the hex location, would perceive and be exposed to that environment. It is

theorized that this approach of using a hexagon tessellated surface could help mitigate the MAUP and UGCoP. Each hex making up the surface is coded with environment variables and will act as bins to store the combined GPS points, by child, resulting in the time spent in each hex area (a proxy for engagement). If a hex is visited by a child's GPS track then that hex is coded as one in which the child has been exposed to and is included in the child's activity space. Separately, the time elapsed in each hex will act as a proxy for engagement in the environment stored as a hexagon variable. Therefore, combining GPS tracks and hex-bins introduces a novel approach in response to the uncertainties and ecological fallacies posed by MAUP and UGCoP (Gilliland & Olson, 2013; Gilliland et al., 2011).
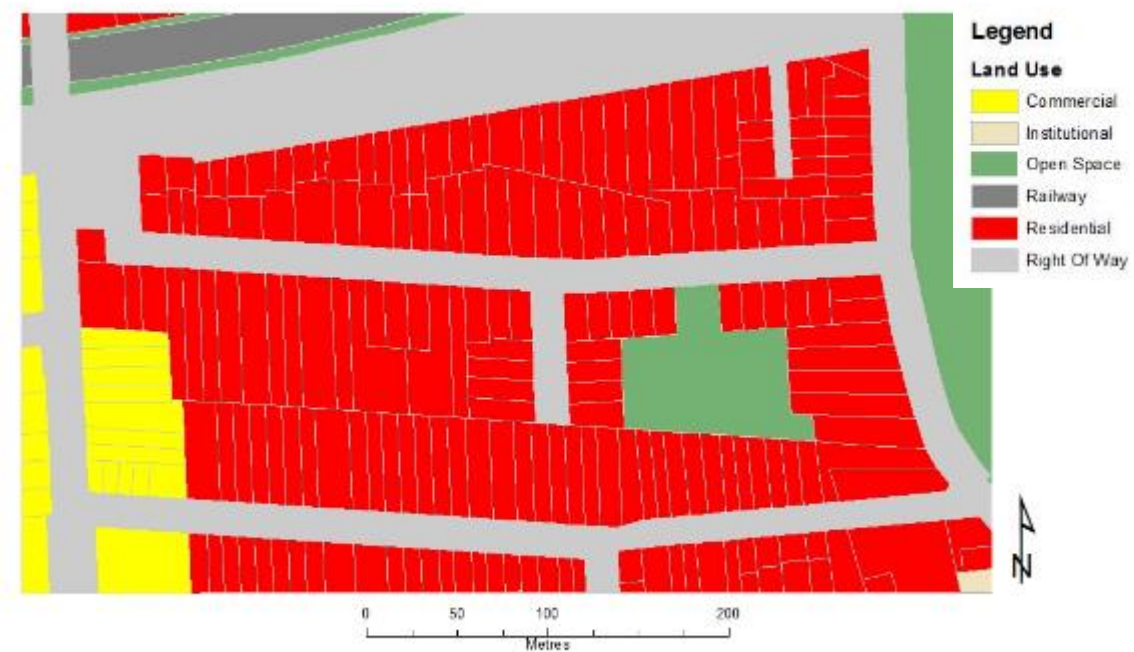
Figure 4.8: Built environment variable map (Land Use)
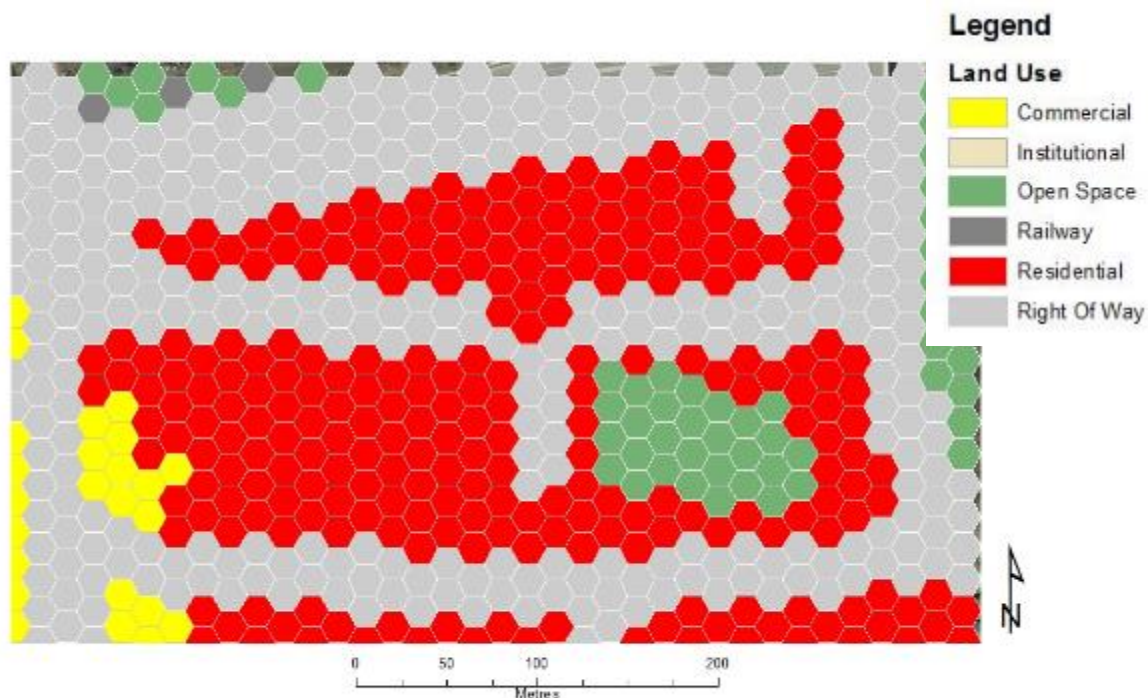
Figure 4.9: Hex-bin environment variables (Land Use)

The calculation of time spent in a hexagon was performed in ArcGIS 10.x where the GPS
coordinates were overlaid with the hexagon surface (Point-in-Polygon) and the
corresponding unique hexagon identifiers were transferred to each GPS coordinate in a
one-to-many table join (hex-to-points) as shown in Figures 4.10 and 4.11. The hexagons
visited by each child participant is now available as a proxy for exposure (Figure 4.11)
while the duration of time spent in each hexagon will be the proxy for engagement
(Figures 4.12 to 4.14). Both the exposure and the engagement metrics will act as the
outcome variables in the statistical analysis.
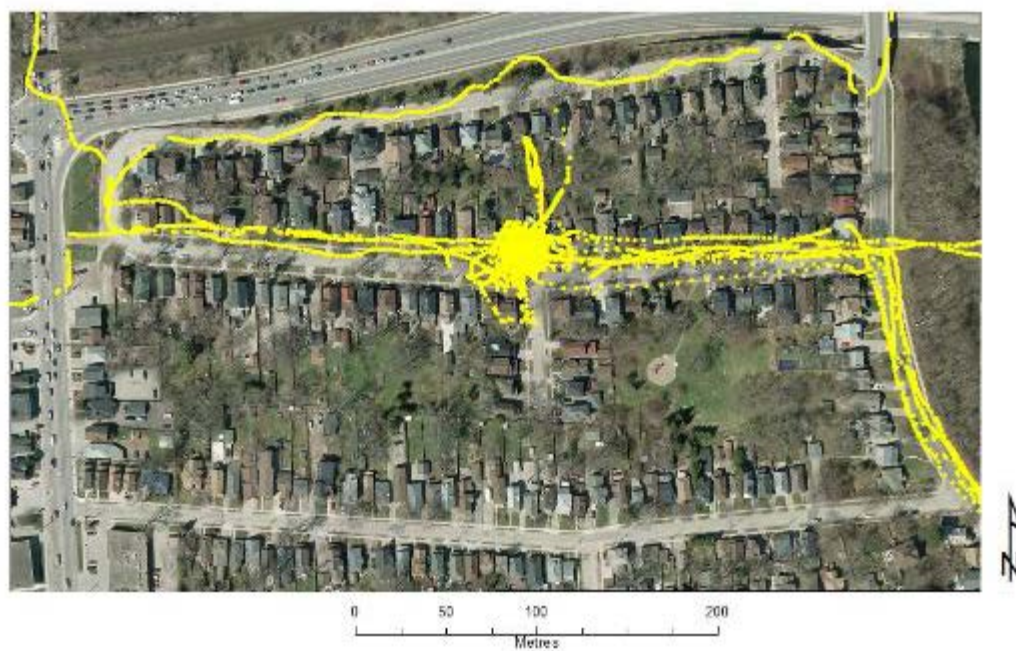
Figure 4.10: Raw GPS tracks
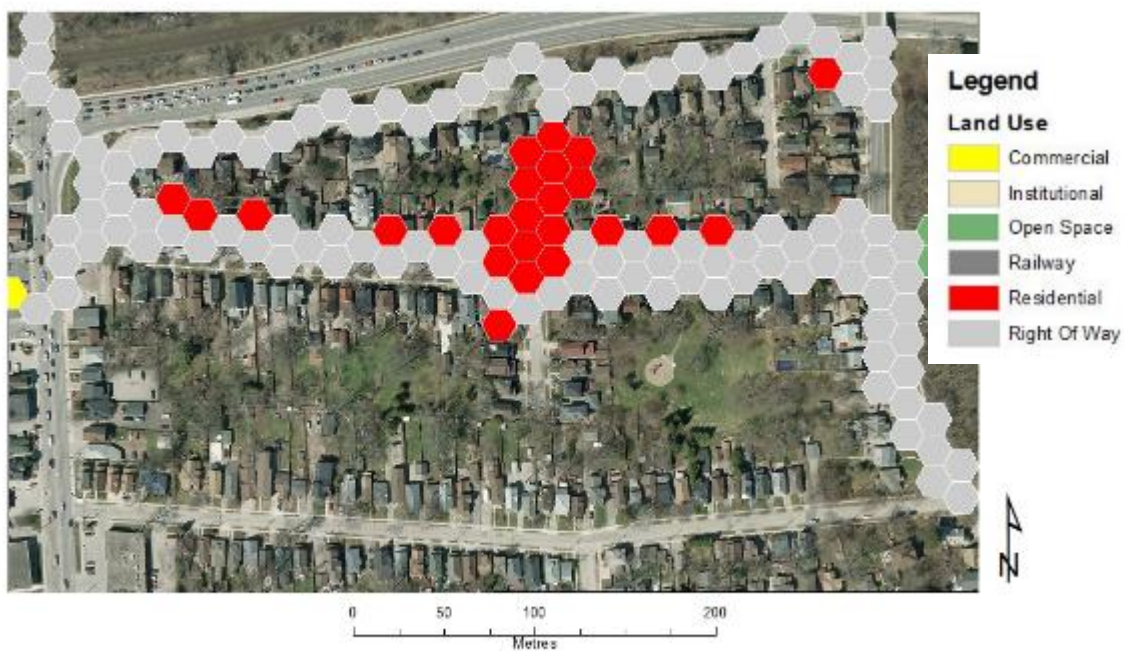
Figure 4.11: Exposure to Land Uses by hex-bin (GPS point in hex)

Figure 4.13: Exposure - time spent in each hex-bin (hotspot)
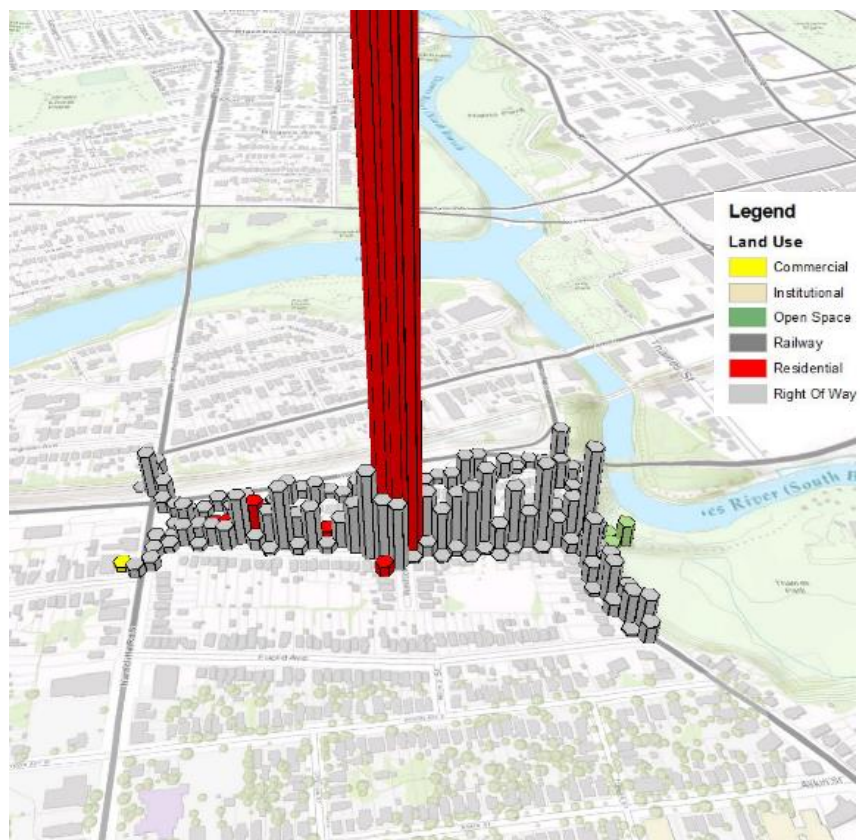
Figure 4.12: Engagement as time spent in hex-bin by land-use (3D view)

4.2.3   Random forest model

The random forest model (Breiman, 2001) is an *ensemble* method of classification which constructs a number of decision trees from a set of random observations from a training dataset. A training dataset is one where the variables are attributed and the final class label has been identified as discussed later in this chapter. As the name suggests, many decision trees are generated creating a 'forest' of decision trees. A decision tree is a type of data classifier as shown in Figure 4.14. Each decision tree is represented by 'branches', connected at 'split nodes', terminating at 'leaves'. An individual leaf of a decision tree represents a class label that is assigned to a data point. To arrive at a class label, the data point 'travels' through the branches going from one branch to another at splits nodes. At the split nodes the data point variable is examined and will move to one branch if the variable value is less than that of the split value, or to the other branch if the data point variable value is greater than the split value. The branches represents the combination of variables/steps that lead to the correct classification of that data point. Many decision trees are created (a forest of them) and the algorithm learns as the forest is being grown so that the misclassification errors are kept to a minimum. The trained model can then us used on the entire data set (see Figure 4.15).

Once the random forest model is built (grown), the entire dataset, one observation at a time can be passed through each decision tree in the forest. An individual observation will pass through every tree in the forest, its corresponding variable checked at each node until the leaf is reached. The class label generated at the leaf is called a vote (i.e. in a forest of 500 trees there will be 500 class label votes per observation). The votes are then tallied and the majority class label is assigned to that observation. The process is repeated for the *n* count of observations (see Figure 4.14)
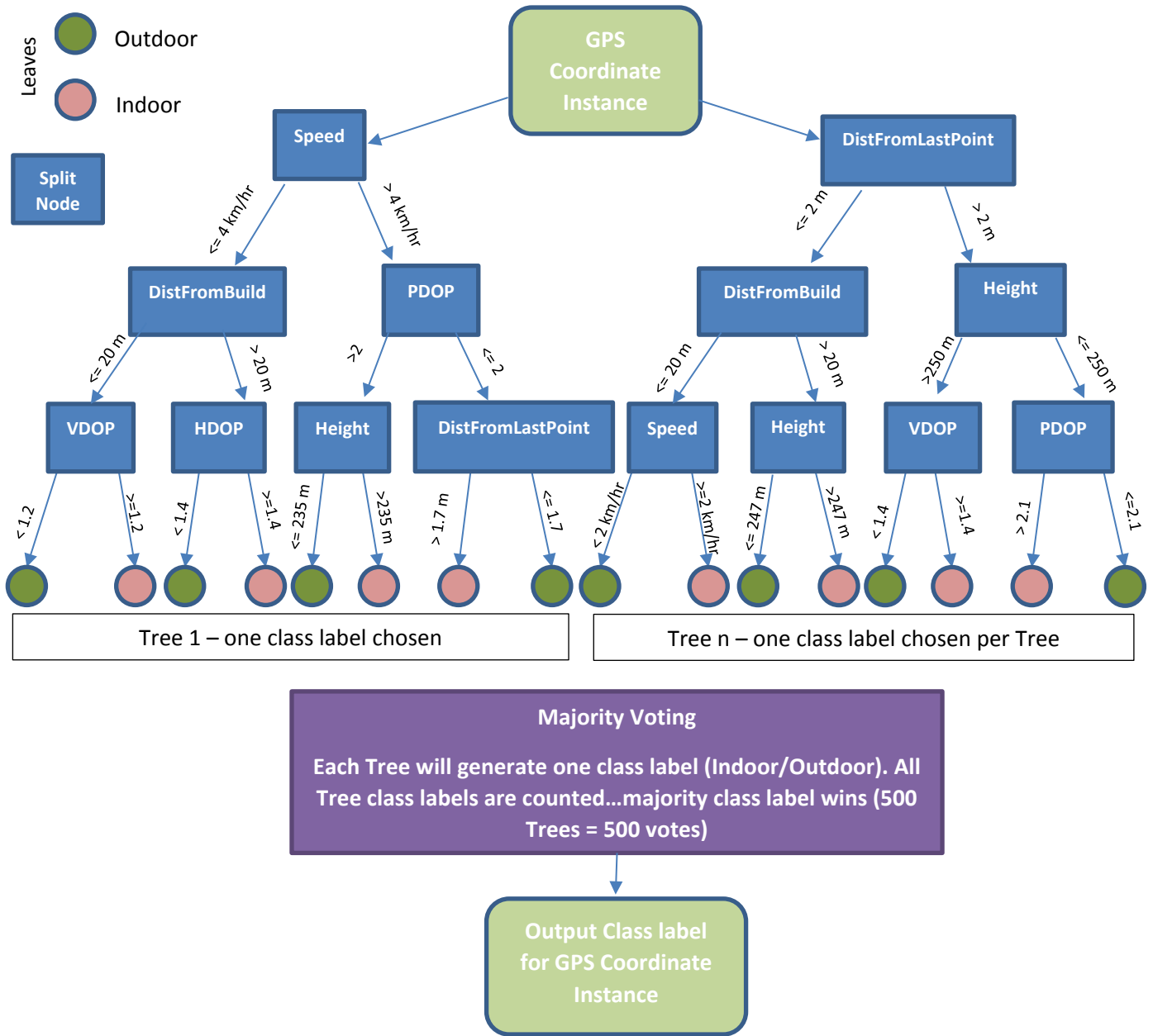
Figure 4.14: Example random forest decision trees

Begin

For each new tree (*i*)

i = **500**

Randomly choose (Bootstrap) training subset (*n*)
of GPS Points

$n = {}^{2}/_{3}$ of training sample used in the creation for each tree, while ${}^{1}/_{3}$ is held back out-of-bag (OOB) so that classification bias can be tested while the forest is grown resulting in OOB Prediction Error Estimate.

Randomly choose variable subset to be used at a split (e.g. PDOP, Speed, and DistFromBuild)

Gini impurity Index used to select the best variable to use at a split – minimizing the model's overall variance.

For each chosen variable

Choose best split value
(Gini Index)

The will always be only two branches at each split.

If at terminal node

No

Yes

Calculate OOB Prediction Error Estimate

End

Figure 4.15: Random forest algorithm flow chart

4.2.4   Out-of-bag (OOB) error estimate

The Out-of-Bag (OOB) error estimate is generated using a random one third of the training sample that is not used in the creation of the forest. In random forests, there is no need for cross-validation tests or any additional tests to get an unbiased estimate of a test set error. Each separate tree is constructed using a different bootstrap sample taken from two thirds of the original data. An entire one-third of the observations are left out of the bootstrap sampling and not used in the construction of any tree. These observations are passed through the completed model to see how well the model was made. The lower the OOB error the more accurate the model. The confusion matrix, shown in Figure 4.16 illustrates how the overall model accuracy and OOB accuracy rates are calculated. The OOB error rate is equal to the *FP+FN/Total OOB sample size* while the model accuracy is calculated by *(TN+TP)/ Total OOB Sample* size. A Confusion Matrix is a method of appraising binary classification procedures, in this case the binary classification of indoor or outdoor.

		Actual (Reference)	
		False	**True**
Predicted	**False**	**True Negative (TN):** Predicted No, and Reference was No.	**False Negative (FN):** Prediction was No, and Reference was Yes.
	True	**False Positive (FP):** Predicted Yes, and Reference was No.	**True Positive (TP):** Prediction was Yes, and the Reference was Yes.
(OOB Sample Accuracy)		TN / (TN + FP) % model correctly predicts No.	TP / (FN + TP) % model correctly predicts Yes.

Figure 4.16: Training data confusion matrix

Generally, the greater number of trees that make up the random forest offer better classification results and hence, more reliable estimates from the out-of-bag (OOB) predictions. You can use OOB error rate to determine the size of the forest (number of trees) as shown in the random forest error vs forest size plot in Figure 4.17. The plot shows that as you increase the size of the forest the improvement rate of the model decreases, as the number of trees increases to a point where in the benefit in prediction performance of adding additional trees levels off, in the example in Figure 4.17 at approximately 280 trees.

All random forest modelling was performed in R (R Core Team, 2018) within RStudio (RStudio Team, 2015). The R 'arcgisbinding' bridge was used to control the reading and writing of spatial data to and from R and the ArcGIS (ESRI, 2018) GIS geodatabases. Additionally, the R Caret package (Kuhn, 2008) used in the creation of the test data sample confusion matrices.
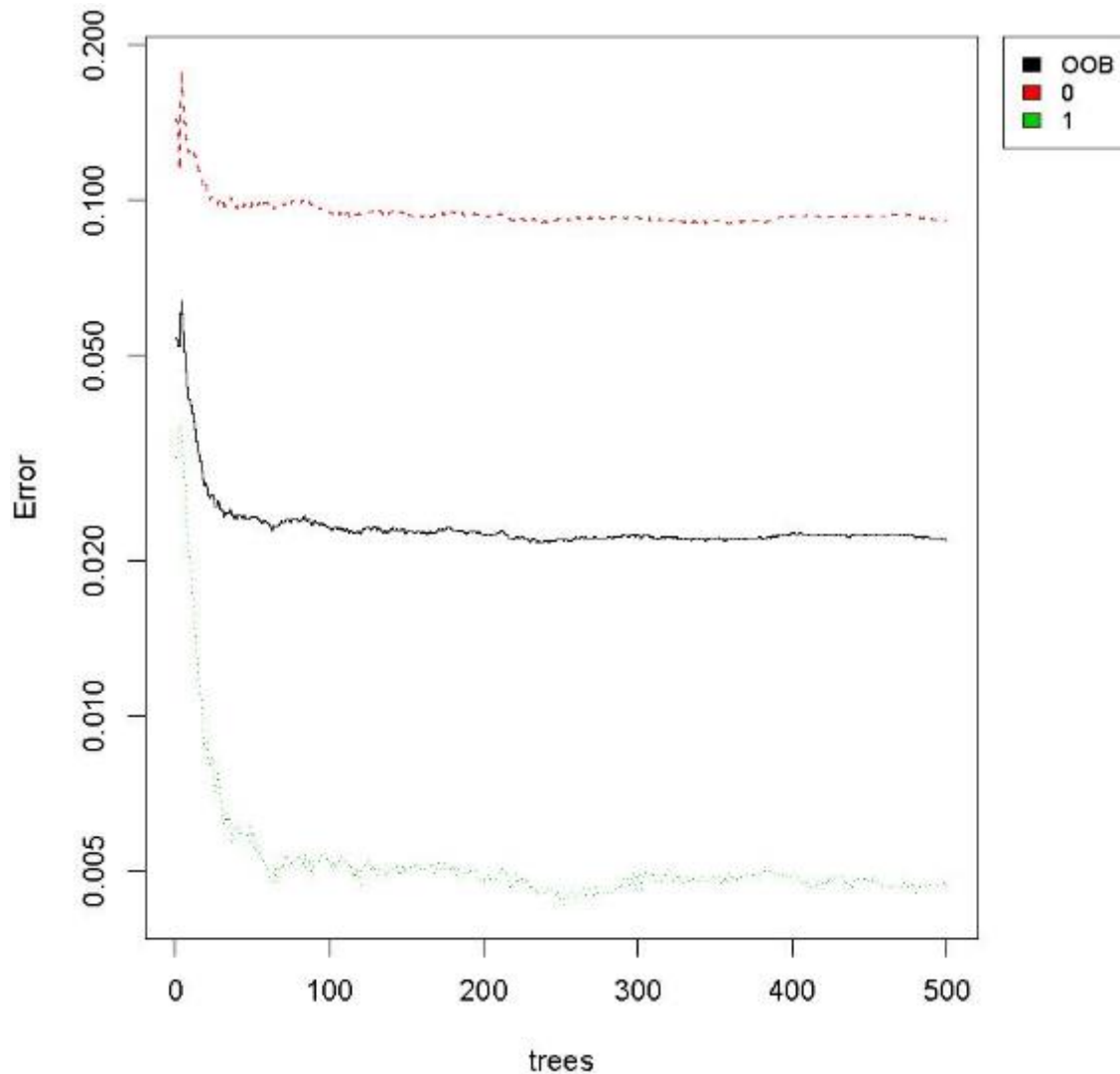
Figure 4.17: Random forest error vs forest size plot

4.2.4.1   Split nodes and Variables

In this study the tree splits nodes will use a set of variables created by the GPS receiver and a set of variables created specifically to improve the model's performance. The GPS signal quality variables, speed, and height from the NMEA GPS sentence will be used. In addition two custom variables; 1) distance between successive GPS points. And 2) distance from blocking structures (buildings).

Standard GPS technology generates an absolute position, based on pseudorange measurements. Each pseudorange corresponds to the distance between the receiver and each GPS space vehicle (SV). The coordinate position (X,Y,Z,T) of the receiver is determined by more than four pairs of pseudorange measurements and their corresponding GPS SV positions. Embedded in the pseudorange signal is information on ionospheric and tropospheric signal delays, clock errors, position and health of the SV in space (Kojima et al., 2012).

The GPS variables used in this study are PDOP (Positional Dilution of Precision), HDOP (Horizontal Dilution of Precision), VDOP (Vertical Dilution of Precision), speed, and height.

Positional Dilution of Precision is given as:

$$PDOP = \frac{\sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2}}{\sigma} \qquad (4.2)$$

Where σ_E^2, σ_N^2, and σ_U^2 are the variances of the east, north, and vertical parts of the receivers position estimate, σ is the standard deviation of the pseudorange measurement error (sent by the space vehicle) and model error which is assumed to be constant for the epoch that the GPS receiver is being used (Langley, 1999).

The Horizontal Dilution of Precision is similar to the PDOP, except for the exclusion of the vertical part of the receiver's position estimate and is written as:

$$HDOP = \frac{\sqrt{\sigma_E^2 + \sigma_N^2}}{\sigma} \qquad (4.3)$$

The Vertical Dilution of Precision uses only the vertical part of the receiver's position estimate and is given as:

$$VDOP = \frac{\sqrt{\sigma_U^2}}{\sigma} \qquad (4.4)$$

Vertical dilution will always be larger than HDOP because, in order to get a more accurate measurement of height, the receiver would need to use pseudorange measurement from all directions, not just those directions above the receiver, but below the receiver as well (Langley, 1999). The NMEA GPS sentence used in random forest model also includes speed (distance/time) and height.

The Gini Impurity (for classification purposes) is tested throughout the random forest model's growth (as the model learns) and measures the variance of how often a randomly chosen observation from a set of random observations would be incorrectly classed if it were given a random class value and is expresses as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.5}$$

where $\hat{p}_{mk}$ is the probability of observations in the mth sample from the kth class (variable) being correct. In our case, the full set of observations comes from 66% of the total GPS training sample per participant/season. Random forest creates the mth bootstrapped sample from $\frac{2}{3}$ of the 66% chosen at random, keeping in mind that a full $\frac{1}{3}$ of the observations are "held back" for the out-of-bag (OOB) error estimate. Each tree is built in a way to keep the *G* low. At the heart of it, the Gini Impurity is a measure of variance. Throughout the creation of the decision tree forest, the variance is summed with each tree's new split. If the addition of a variable for a new split increases the variance in the random forest model, the algorithm will swap out that variable at that split for another variable until the overall variance is reduced. The higher the variance, the more misclassification will exist. Therefore lower values of the Gini Impurity will yield a better classification result which can be tracked using a variable importance plot (see example in Figure 4.18). The plot shows the variables that play a larger role than others in a Random forest classification model. The model will rely more on those variables that increase classification accuracy than those that do not. In this example the *DistFromBuild* and *HEIGHT* variables play a larger role in accurately classifying (reducing Gini Impurity), than do the *DistFromLastPoint* and *PDOP* variables..
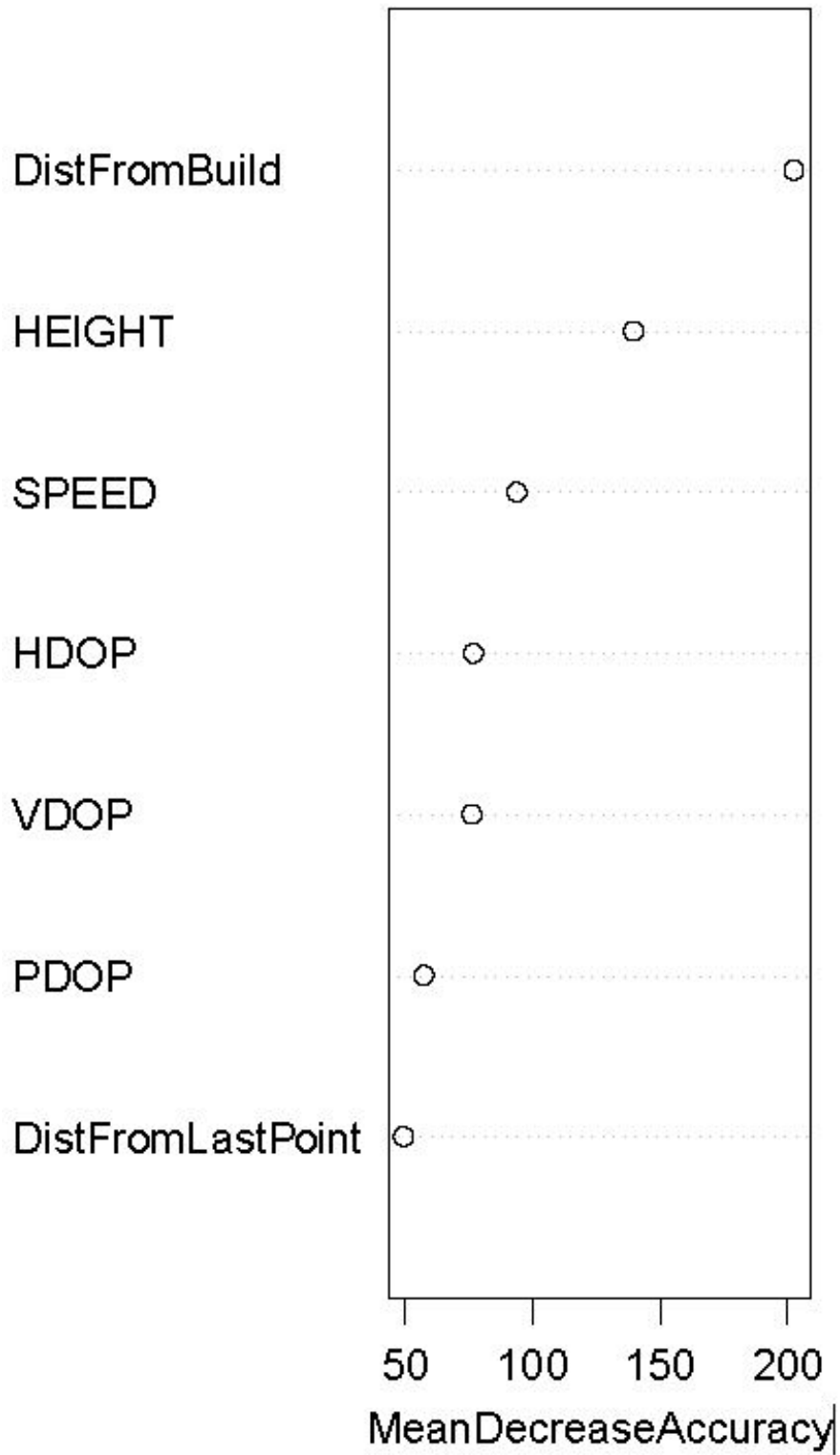
Figure 4.16: Example variable importance plot

4.2.4.2   Creating the training sample

There are set epochs of time during each day of the study where the location of the participants in the project is known. These times are primarily when the children are attending school. Each school schedule was logged and joined to each GPS coordinate so that during school hours it is known whether the participant is inside the school (in the classroom) or outside during recess time and other breaks. Each participant activity diary was then used to verify whether the child was inside or outside for recess. The GPS units meanwhile continue to log data points which, for the most part, fall spatially within the school building during class time, while some are generated outside of the building footprint. The opposite is true when the children are spending time outdoors. Most logged GPS points appear outside of the building while some appear inside, see illustration in Figure 4.19. These scheduled times, and their GPS coordinates, by participant by school day will act as the training dataset for the random forest model. To make the argument that the participants' GPS school-time dataset, typically about 7 hours per day, represented a valid snapshot of their indoor/outdoor activity, it was determined that only days that children spent at the school were included in the training sample. During the school day, only four scheduled time blocks were used for training: AM In School, AM Recess, PM in School, and PM Recess). A proximity measure was applied to each of the
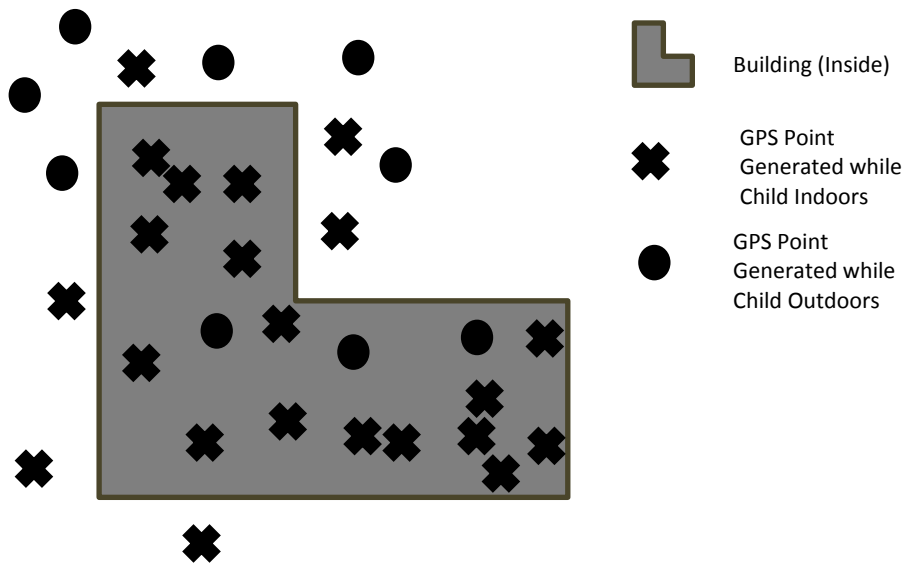
Figure 4.17: Misclassification of GPS points

participant GPS tracks during each day and those tracks whose mean distance from the school building was greater than 500m were not included in the scheduled school time block of the training sample. The participant's daily activity diary were examined to validate the child as being indoors or outdoors during recess. The training sample was bolstered by the inclusion of GPS points identified as non-active (vehicular) travel from routes created by the *SphereLab Tool* (Thierry et al., 2013) . A route was flagged as non-active travel when the speed cut-off matched the mean speed children (13 years and under) ride a bicycle (13.3 km/hr.). The vehicle trips were coded as indoors and all other trips as outdoors. The addition of the children's travel behaviour GPS points was critical to include in the training sample because: 1) the totality of a child's activity during the day does not always match with what children experience during school time on school property; and 2) some training samples sizes were too small (too small a sample in outdoor time, or indoor time, or both), thus weakening the random forest model. The distance between successive GPS points was also created as an additional custom variable to address, in part, the errors in the GPS calculated *Speed* variable. Occasionally, a GPS coordinate showed a speed of zero, which would indicate no movement, but the proceeding and successive points were shown to be not spatially coincident. Upon further examination this behaviour seemingly occurred while the GPS was under the influence of a blocking structure (building).

All the pre-preprocessing was performed using a custom stand-alone python application (Python Software Foundation, 2018) given in Appendix D.

4.2.5   Processing the GPS points

The processing of the GPS points in R (R Core Team, 2018) is straightforward. In this study the GPS data were stored in a ArcGIS (ESRI, 2018) Geodatabase which was connected to R using the R-ArcGIS Bridge 'arcgisbinding' Package (ESRI, 2018). RStudio (RStudio Team, 2015) was used as the development environment. The R script which trains, builds, and implements the random forest model is given in Appendix E. As the final step, the outdoor GPS points were spatially overlaid with the hex surface so that a link be made between GPS point and the environment (stored per hex-bin). Where exposure is represented by number and aggregated spatial extent of the hexes visited and

engagement is represented by the time spent in each hex. The overall processes is shown Figure 4.20.
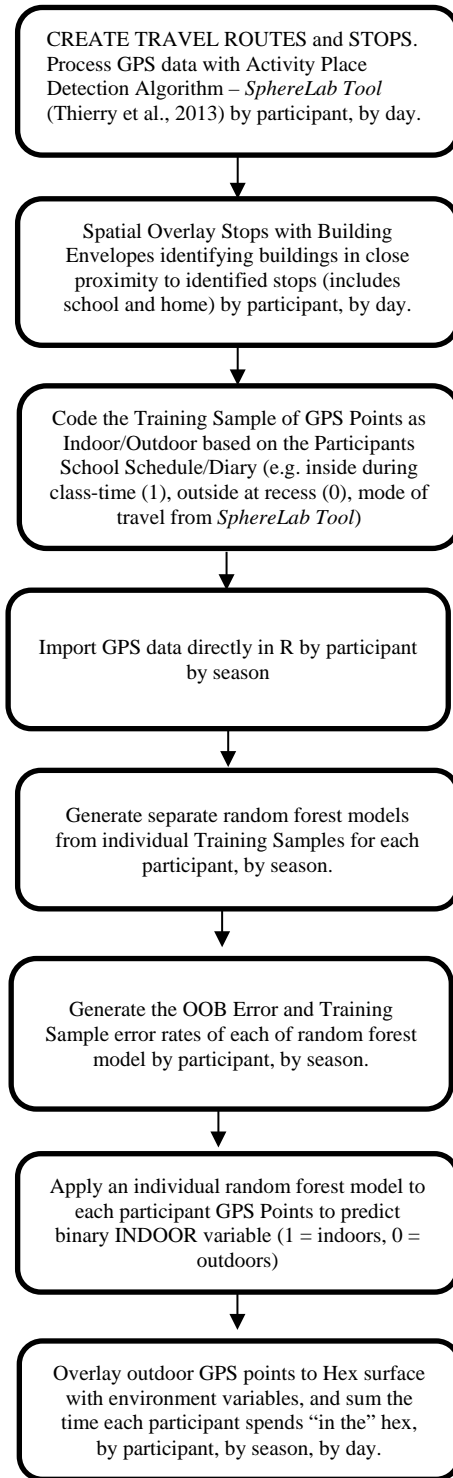
Figure 4.18: Process to code and classify GPS points

4.2.6   Measuring exposure and engagement

All descriptive and inferential statistics were generated using SPSS 25 (IBM Corp., 2017) to examine where and for how long children engage outdoors and how *individual-level*, and *neighbourhood-level* socioeconomic and environmental factors, as per the socio-ecological framework, could aid or hinder this behavior. Both the measures of exposure to, and engagement at, these outdoor locations are used as dependent variables in this study.

The total outdoor exposure is defined as direct contact with outdoor environmental features. It is operationalized by the sum of all the corresponding hex-bins coincident to the child's outdoor GPS points (i.e. hex-bins in direct contact with the GPS tracks) to represent the child's total outdoor "activity space". The Total Outdoor Exposure by environmental feature is measured as a proportion of the child's total "activity space":

$$Exp = ThEnv \, / \, Th \tag{4.6}$$

where *Exp* is equal to the proportion of total hex-bins by specific environmental feature (*ThEnv*) divided by the total outdoor 'activity space' (*Th*).

In this study, engagement outdoors (i.e. total time spent outside per day at parks) will be operationalized by summing the outdoor time spent by the child in each hex-bin. The total engagement at specific environmental features outdoors is measured by:

$$Eng = TtEnv \, / \, Tt \tag{4.7}$$

where *Eng* is the proportion of outdoor activity time and *TtEnv* is the total time a child spends at particular environmental feature type, and *Tt* is the total time the child spends outdoors.

As outlined in Chapter 1 (Figure 1.2), the socio-ecological framework (Bronfenbrenner, 1979; Sallis et al., 2006; Stokols, 1992) was used to guide our understanding of the multi-level influences that recognizes that there are many types of influence on children's behaviours and health outcomes (Sallis et al., 2006; Stokols, 1992). In this analysis, the *Individual-level* factors, related most directly to the individual, and those which are theorized

to influence outdoor exposure and engagement are: age (9-14 years), gender (females / males), and visible minority classification. The *individual-level* factors were taken from the STEAM survey instrument. The *Household-level* factors under examination include: the household structure (dual-parent/ lone-parent household) from the STEAM survey instrument, and socio-economic status (as represented by the median household income of the census dissemination area in which the child's home is located). The *Neighbourhood-level* factors measured at the hex-bin level, are: land use types, parks, street tree counts derived from spatial data published by the City of London (2010-2013) and green space (City of London, 2008). The green space was generated using a Normalized Difference Vegetation Index (NDVI) which commonly used as a means of classifying greenness from infrared aerial imagery. The City of London (2008) infrared aerial photography taken in August 2008 at 30 cm resolution was used to generate the NDVI layer which was then classified to derive two green attributes: heavily forested areas and areas of vegetation and turf.

The Shapiro-Wilk test of normality was used on all variables to test if they come from a normally distributed sample. The Wilcoxon-Mann-Whitney Test was used to calculate the bivariate relationships between the dependent variable (exposure or engagement) and the categorical independent variables. ANOVA and T-Tests were used to compare the mean exposure and engagement by the *individual*, *household*, and *neighbourhood-level* variables.

4.3   Results

4.3.1   GPS Accuracy and Precision

Accuracy refers to the closeness of a measured value to a standard or known value and can be measured using the Root Mean Square Error (RMSE). In this study a VGPS-900 GPS device was carefully placed above a horizontal survey monument with a confirmed positional accuracy of 1 millimetre. The Root Mean Square Error of the VisionTac VGPS-900 GPS device is over 1/2 hour of operation is 0.373m in the X (UTM Easting) and 2.310m in the Y (UTM Northing). The mean PDOP was 1.586 and the HDOP was 0.862. The test is considered the best case scenario for GPS data collection where there is a cloudless clear sky view with no obstructions above 20º from the horizon.

At the 10 second point into the test, the GPS started generating GPS measures using the WAAS differential signal. Positional errors in the Y were corrected as the test progressed and at the five minute mark of the test the Root Mean Square Error was lower for both X (RMSE = 0.314m) and Y(RMSE = 1.574m). The mean PDOP was slightly higher at 1.610 and the HDOP was lower at 0.852. The PDOP includes elevation in its calculation so a higher value for this metric was not unexpected.

Precision refers to the closeness of all the measured values to each other. It was calculated by measuring how far the GPS coordinates deviate from their total mean centre point. The precision of the sample of GPS points is 0.857 m, meaning that the majority of the GPS points measured are within less than a metre of the mean centre. The mean centre is 1.966m from the survey monument.

The manufacturer's reported Circular Error Probability (CEP) of the VGPS-900 GPS receiver during optimal conditions with full sky view, and wide augmentation assisted system (WAAS) enabled (Differential GPS -DGPS), is 1.5m CEP (30%-50%) p=0.05, and 2.5m CEP (95%) p=0.05. In this study, the test revealed that the GPS receiver had an accuracy at 1.5m CEP (43%) p=0.05 which is within the manufacturer's published accuracy claims. However, when testing at 2.5 metres it was found that in our 2.5m CEP (79%) p=0.05 results fell well short of the expected 95%. Therefore, the results show that VisionTac GPS, in optimal conditions, generate 43% of the coordinates within 1.5 m of the true location, and only 79% of the coordinates within 2.5 m of the true location, 19 times out of 20.

In the STEAM study, on average, 34.9% of the GPS points were generated with differential GPS quality of PDOP < 2.5, so that with optimal conditions, on average, 34.9% of the points were within 5m of the true location 19 times out of 20. It is expected that the accuracy of the units will degrade when in close proximity to blocking structures, making these numbers are unrealistic.

The map in Figure 4.19 illustrates the actual cluster of GPS points generated around the survey monument. The points in red indicate the most displacement and occurred in the first 5 minutes of the test.
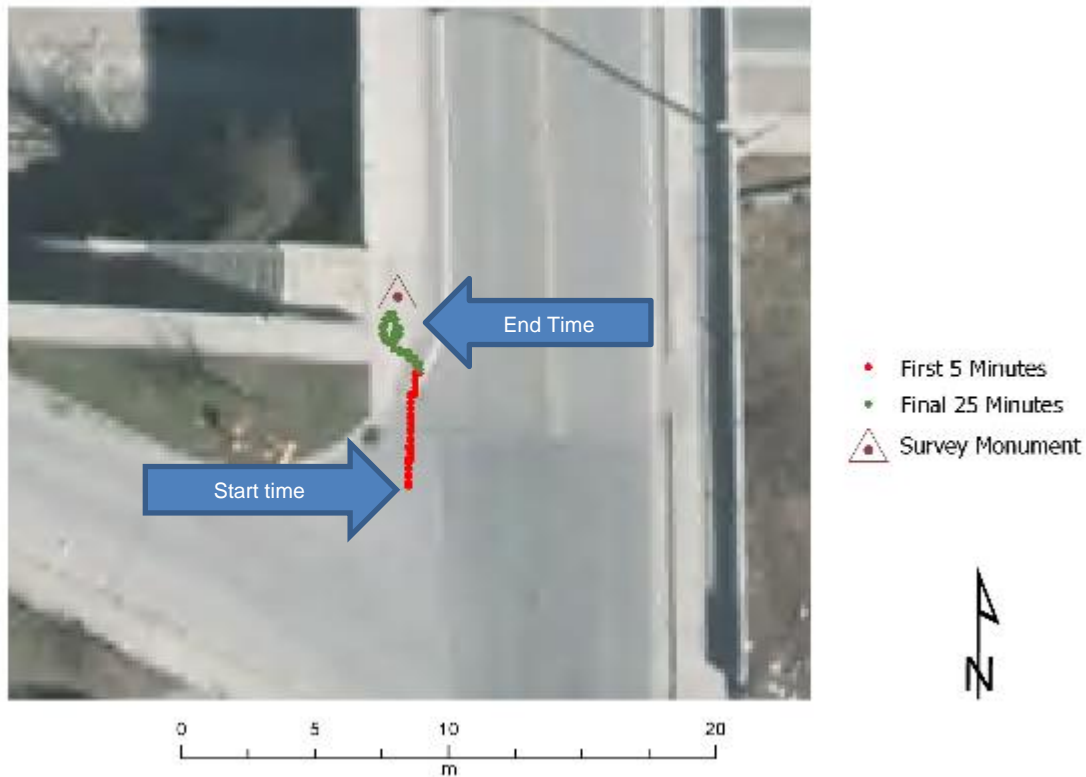
Figure 4.19: Map of GPS test at the survey monument

4.3.2 Random forest outdoor classification model

In this study only STEAM 1 participants were used in this study, and a separate random forest model was created, for the most part, for each participant by season. The mean training sample size by season (reported in Table 4.1) tends to be similar, as do the errors rates. The mean weekly training sample size equaled 16,544 GPS coordinates for the spring season representing a mean of 4.6 hours of training data per week, while for winter the training sample size was 16,702 GPS coordinates representing a similar amount of time in the week. The mean OOB Error rate of 0.01 was the same for both seasons and the test data accuracy was similarly high with values for spring at 98.9% and 99.9% for winter. Written another way, on average the random forest models identified GPS points as indoor or outdoor correctly 98.9% and 99.9% of the time on average. The variable

importance plots, an example plot of a single model shown earlier in Figure 4.16, show how much the model would decrease in accuracy if any variable listed were excluded from the model. A separate variable importance plot was created for each model. In the example (in Figure 4.16) and for a vast majority of the models the splits at variables *distance to building, Height* and *Speed* were the most important in the classifications. Somewhat surprisingly the dilution of precision measures, for the most part, were less important in categorizing indoor/outdoor. It was also found early in the research that if the training sample only was created using scheduled school times and not including the travel modes there was a 10% average drop in the overall accuracy of the classes from the random forest model. The reason for this was identified by the travel behaviour of participants while not at school. The vehicle travel was not modelled in the training data and therefore resulted in obvious vehicular travel being classified as outdoor activity. The remedy was to include GPS points identified as travel routes by the *SphereLab Tool* (Thierry et al., 2013) and coded as vehicle travel (indoor) beforehand in the training samples.

Table 4.1: GPS classification

	n	GPS Observations			Training Sample Size		OOB Error		Test Sample Size		Test Accuracy	
	n	**n**	**$\bar{X}$**	**sd**	**$\bar{X}$**	**sd**	**$\bar{X}$**	**sd**	**$\bar{X}$**	**sd**	**$\bar{X}$**	**sd**
Winter	72	11,768,265	154,845	46,807	16,702	9873	0.01	0.009	7405	4377	0.999	0.009
Spring	62	10,357,609	143,855	60,922	16,544	9671	0.01	0.005	7335	4287	0.989	0.004

4.3.3 Exposure outdoors

As discussed earlier in this chapter, the GPS tracks for the children were spatially overlaid with a hex-bin surface which encoded the neighbourhood-level environment variables (e.g. green space, and built environment). The result of the overlay is a spatial cluster of all the hex-bins visited by the child which is considered the child's *activity space*.

Using hex-bins to estimate exposure allows for a closer approximation of activity space than most traditional methods and is expressed as an encounter with the outdoors and an encounter with any natural and built environment features visited outdoors. It is operationalized as a proportion of the total activity space (area). As seen in Table 4.2, the mean activity space for all participants in the study is 37.6 hectares and varied widely among participants (*sd* =29.9). Furthermore, the findings suggest that children are much less mobile on weekends compared to weekdays as the weekend activity space (28.48 ha, *sd*=25.72) is just under half of what it is on weekdays. The activity space is just under half of what it is on weekdays as seen in Table 4.3

Table 4.2: Total outdoor activity space by *individual-level* variables

	Total Outdoor Activity Space (area per ha)	
	$\bar{X}$	sd
All	37.62	29.94
Boys	32.36	30.63
Girls	39.61	29.53
Age		
10 or under	44.73	34.74
11	38.01	29.02
12	34.92	24.66
13	42.32	35.38
14	41.19	33.58
Weekday	46.20	30.74
Weekend	28.48	25.72

*Only full days during each study included (First and final days removed)

4.3.4   Engagement outdoors

On average, children in this study, spent 18.5% of their recorded time outdoors which translates to an average of 81 minutes (1 hour and 21) of time spent outdoors per day (Table 4.3). For total outdoor engagement, children are outside for a larger portion of the day in winter than in spring on average, both for weekend and weekdays, however the results were not significant (*p*=0.158).

Table 4.3: Engagement outdoors by day type and season (proportion of time)

	N	$\bar{X}$	sd
Winter Weekend	64	21.2585	25.56077
Winter Weekday	72	21.1090	21.57640
Spring Weekend	51	16.8013	19.23718
Spring Weekday	62	14.1375	15.61773
Total	249	18.5292	21.05109

The results using the variables from the *Individual-level* analysis indicate that there are no significant differences between genders or ages in the amount of time children spend outdoors. This study showed that Children who are visible minorities spend more time outdoors in spring than Caucasian children, but not significantly so. Additionally, it appears that children, regardless of age or gender spent similar amounts of time across the various land uses. On weekdays there is no significant difference between the amounts of time children are spend outdoors at different land uses; in other words, children are spending their time outside similarly when it comes to the proportion of time spent at each land use.

The Household-level analysis shows that only socio-economic status (median household income) showed as a statistically significant factor when predicting outdoor activity with children living in wealthier neighbourhoods being more likely to spend more time outdoors.

The findings for the *Neighbourhood-level* analysis indicate that on average, children living in the suburbs are more likely to spend more time outdoors than children living in more urban settings. Additionally, those children who only spent time outdoors in less varied land uses (such as residential and institutional areas) than those children who spent time in a wider array of land uses spent overall less time outdoors than their counterparts.

The results indicate that children spend more time outdoors when they are engaged in varied land uses. In spring, not surprisingly, children are most likely to spend time outdoors in parks than in the wintertime. During winter, though children spent more time outside in residential areas than they do in the springtime.

4.3.5   Outdoor engagement vs exposure

Identifying how long children spend at a geographic setting within the activity space cannot be measured by exposure. A more realistic measure is the based on the amount of time children spend engaging in an environment, rather than being simply exposed to it. When comparing the proportion of time spent outdoors (engagement) and the proportion of activity space traversed outdoors (exposure) we see, in Figure 4.20, that children engage in land uses in a way not reported with exposure. Case in point, commercial areas are the second largest exposure proportion for children on weekdays, but when looking at
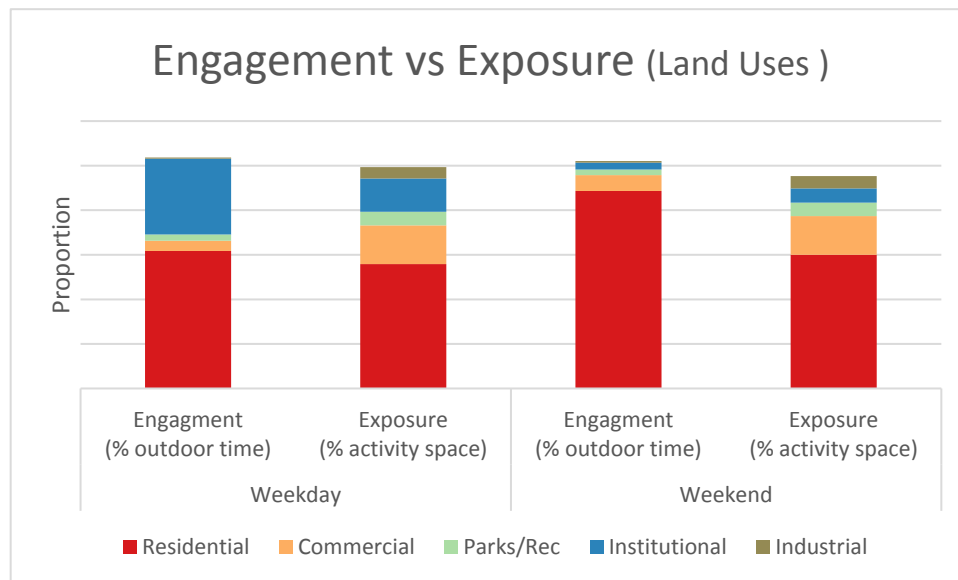
Figure 4.20: Engagement vs exposure proportions by land use

time spent (i.e. exposure), it is marginal activity. The residential exposure is similar on both weekdays and weekends where just over half the activity space outdoors being spent there, but the exposure tells a more complete story. The children are almost spending

their entire outdoor time in residential areas on weekends the magnitude of which is not represented with exposure. The same can be said for institutional (schools included), on weekdays the engagement is high, while the exposure is relatively low.

When comparing children's engagement at green space versus exposure to green space, as seen in Figure 4.21, children are heavily exposed to more areas with street trees, but spend proportionally less time in these areas (engagement). Conversely, the proportion of time spent engaged in green vegetation and turfed areas is higher than the exposure metric. One weekends and weekdays, the children are similarly engaged in and exposed to green areas, and less than one tenth of 1% of time/activity space is spent in forested areas.
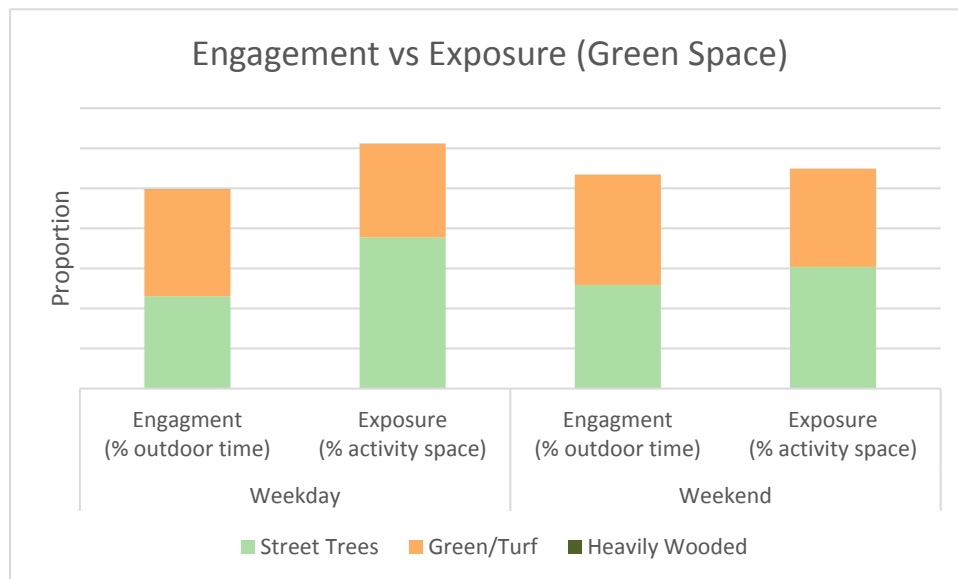
Figure 4.21: Engagement vs exposure proportions by green space

4.4   Discussion

Dramatic increases in children's sedentary behaviour indoors is a growing concern for health researchers. An objective way to identify and quantify outdoor activity has yet to be fully realized (Wang et al., 2018); and therefore, it was the intention of this study to describe and test a methodological breakthrough for overcoming this problem. This study

was divided into two parts: 1) it described a method made up of a combination of three novel approaches that in sequence practically measure, classify, categorize, and map children's spatial behaviours outdoors and links the environmental features in their neighbourhoods to measure exposure to and engagement in the outdoors; and 2) it identified statistically significant metrics at the *neighbourhood-level* (e.g. green space, land use types), and at the *household-level* (e.g. income), and identified interesting trends at the *individual-level* all in an effort to identify factors from multiple levels of the socio-ecological model that are associated with and can be used to predict outdoor exposure and engagement.

The inclusion of stops generated from *SphereLab Tool* (Thierry et al., 2013) were critical to establish the most important variable in the random forest models: *distance from buildings*. The stops were employed to identify buildings located in proximity to the stops and treated as potential blocking structures or destinations so that a *distance from buildings* variable could be used in the random forest classification. The *distance from building* variable was consistently the most important variable for correctly classifying indoor and outdoor GPS points. It was, also found that relying on only the school scheduled times for the training sample resulted in a 10% average drop in the overall accuracy of the classes from the random forest model without the routes identified by the *SphereLab Tool* beforehand.

4.4.1   Comparison with previous participant surveys studies

Our GPS-based methods revealed that on average children in this study spent 18.5% of their recorded time outdoors translating to 81 minutes (1 hour and 21) per day outdoors. In comparison, a previous study (Matz et al., 2014) used parental report data to estimate that children across urban and rural Canada, aged 5-11 years (n=428), spent 1 hour and 48 minutes outdoors per day. In this Chapter, the results show that participants in this study spent considerably less time outdoors than reported in most other research studies who have relied on participant surveys alone. Milne et al. (2007) found that in a study of Australian children, aged 6-12 years (n=1614), spent between 2 and 3 hours outdoors during daylight hours. Fifty percent of U.S children (n=1822) spent just over 2 hours a

day outside in a study by Kimbro et al. (2011), while Chinese children (n=681) spent 97 minutes a day outside (Guo et al., 2013).

Most of the questionnaires in these studies were completed by parents which inevitably cause recall errors. Apart from recall bias, the estimation of time outdoors was likely to be more difficult for a parent to answer. Compounding the recall bias and errors, these surveys were designed differently and the qualitative descriptions could be imprecise, so it might be unsuitable to compare the empirical results from this study with those from recall based studies.

4.4.2   Comparison with previous GPS studies

The majority of research studies employing GPS to identify outdoor activity use the PALMS protocol. In terms of children's activity outdoors, Klinker et al. (2014) found that 11 to 13 ($\bar{x}$ = 12.4) year old children (n=129) in Copenhagen, Denmark spent a median of 226.5 minutes (3hr 4 m 36s , IQR (175 - 284.5) outdoors. In contrast, Tandon et al. (2013), in a study of pre-school aged children, reported that children spent on average 63 minutes a day outdoors. The participants were younger than in our study and Tandon included a wearable sunlight sensor on the participants to identify time outdoors. A study by Cooper et al. (2010), as part of the PEACH project, indicated that British school aged children (n=1010) spent only 41.7 minutes outdoors. A recent study of British children (n=70) by Pearce et al. (2018) found a median total outdoor time of 80.3 minutes, which most closely matches our results. In their post-processing of the GPS, Pearce filtered the signal-to-noise cut-off at (SNR<=212) which is lower than the standard PALMS tool cutoff (SNR<=250) used by Klinker and colleagues (2014), suggesting that Klinker's study in Denmark over-estimated time spent outdoors.

4.5   Conclusion

Rising public concerns over children's health has led to a growing public and academic interest in gaining a better understanding of the role that the physical environment (built and natural) plays in mitigating or exacerbating health issues (Tillmann et al., 2018). There is a realization in public and academic circles that children are spending less time outdoors than ever before, (Gilliland, 2018; Zorzi & Gagne, 2012) which has led to an

increase in studies exploring the association between time spent outdoors and some physical and mental health outcomes. Yet, much of this research is difficult to decipher due to incomparable, imprecise, or inaccurate methods for assessing environmental exposure (Tillmann et al., 2018). Ultimately, it is anticipated that this study offers a methodological breakthrough for overcoming the problems inherent in GPS indoor and outdoor classification, and in the linking of exposure and engagement variables in a meaningful way.

4.6 References

Amoly, E., Dadvand, P., Forns, J., Lopez-Vicente, M., Basagana, X., Julvez, J., . . . Sunyer, J. (2014). Green and blue spaces and behavioral development in Barcelona schoolchildren: the BREATHE project. *Environ Health Perspect, 122*(12), 1351-1358. doi:10.1289/ehp.1408215

Breiman, L. (2001). Random Forests *Machine Learning*(45), 5-32.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design.* Cambridge, Massachusetts: Harvard University Press.

City of London. (2008). InfraRed 30cm Orthographic Aerial Imagery.

City of London. (2010-2013). *Parcels, buildings, address points, and health facilities GIS files [DVD]*.

Cooper, A. R., Page, A. S., Wheeler, B. W., Hillsdon, M., Griew, P., & Jago, R. (2010). Patterns of GPS measured time outdoors after school and objective physical activity in English children: the PEACH project. *Int J Behav Nutr Phys Act, 7*, 31. doi:10.1186/1479-5868-7-31

Davis, J., & Robinson, G. (2012). A Geographic Model to Assess and Limit Cumulative Ecological Degradation from Marcellus Shale Exploitation in New York, USA. *Ecology and Society, 17*(2). doi:10.5751/es-04822-170225

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Front Public Health, 2*, 36. doi:10.3389/fpubh.2014.00036

ESRI. (2011). ArcGIS Desktop (Version 10.0). Redlands, CA: Environmental Systems Research Institute.

ESRI. (2018). ArcGIS Pro (Version 2.2.0). Redlands, CA: Environmental Systems Research Institute.

Faber Taylor, A., & Kuo, F. E. M. (2011). Could Exposure to Everyday Green Spaces Help Treat ADHD? Evidence from Children's Play Settings. *Applied Psychology: Health and Well-Being, 3*(3), 281-303. doi:10.1111/j.1758-0854.2011.01052.x

Flouri, E., Midouhas, E., & Joshi, H. (2014). The role of urban neighbourhood green space in children's emotional and behavioural resilience. *Journal of Environmental Psychology, 40*, 179-186. doi:10.1016/j.jenvp.2014.06.007

Gilliland, J. (2018). Lawson Foundation Systematic Review.

Gilliland, J., & Olson, S. (2013). Residential Segregation in the Industrializing City: A Closer Look. *Urban Geography, 31*(1), 29-58. doi:10.2747/0272-3638.31.1.29

Gilliland, J., Olson, S., & Gauvreau, D. (2011). Did Segregation Increase as the City Expanded?: The Case of Montreal, 1881-1901. *Social Science History, 35*(4), 465-503. doi:10.1215/01455532-1381823

Gray, C., Gibbons, R., Larouche, R., Sandseter, E. B., Bienenstock, A., Brussoni, M., . . . Tremblay, M. S. (2015). What Is the Relationship between Outdoor Time and Physical Activity, Sedentary Behaviour, and Physical Fitness in Children? A Systematic Review. *Int J Environ Res Public Health, 12*(6), 6455-6474. doi:10.3390/ijerph120606455

Guo, Y., Liu, L. J., Xu, L., Tang, P., Lv, Y. Y., Feng, Y., . . . Jonas, J. B. (2013). Myopic shift and outdoor activity among primary school children: one-year follow-up study in Beijing. *PloS one, 8*, e75260. doi:10.1371/journal.pone.0075260

IBM Corp. (2017). IBM SPSS Statistics for Windows (Version 25.0). Armonk, NY: IBM Corp.

Kestens, Y., Thierry, B., & Chaix, B. (2016). Re-creating daily mobility histories for health research from raw GPS tracks: Validation of a kernel-based algorithm using real-life data. *Health Place, 40*, 29-33. doi:10.1016/j.healthplace.2016.04.004

Kimbro, R. T., Brooks-Gunn, J., & McLanahan, S. (2011). Young children in urban areas: Links among neighborhood characteristics, weight status, outdoor play, and television watching. *Social Science and Medicine, 72*, 668-676. doi:10.1016/j.socscimed.2010.12.015

Klinker, C. D., Schipperijn, J., Kerr, J., Ersbøll, A. K., & Troelsen, J. (2014). Context-specific outdoor time and physical activity among school-children across gender and age: using accelerometers and GPS to advance methods. *Front Public Health, 2*. doi:10.3389/fpubh.2014.00020

Kojima, Y., Suzuki, N., Hattori, Y., & Teramoto, E. (2012). Proposal for a new localisation method using tightly coupled integration based on a precise

estimation of trajectory from GPS Doppler. *Vehicle System Dynamics, 50*(6), 987-1000. doi:10.1080/00423114.2011.602697

Kuhn, M. (2008). Caret package. *Journal of Statistical Software, 28*(5).

Kwan, M. (2012b). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers, 102*(5).

Langley, R. (1999). Dilution of Precision. *GPS World*(May), 52-59.

Loebach, J., & Gilliland, J. A. (2014). Free Range Kids? Using GPS-Derived Activity Spaces to Examine Children's Neighborhood Activity and Mobility. *Environment and Behavior, 48*(3), 421-453. doi:10.1177/0013916514543177

Matz, C. J., Stieb, D. M., Davis, K., Egyed, M., Rose, A., & Chou, B. (2014). Effects of age, season, gender and urban–rural status on time-activity: Canadian Human Activity Pattern Survey 2 (CHAPS 2). *Int J Environ Res Public Health, 11*. doi:10.3390/ijerph110202108

McCracken, D. S., Allen, D. A., & Gow, A. J. (2016). Associations between urban greenspace and health-related quality of life in children. *Prev Med Rep, 3*, 211-221. doi:10.1016/j.pmedr.2016.01.013

Milne, E., Simpson, J. A., Johnston, R., Giles-Corti, B., & English, D. R. (2007). Time spent outdoors at midday and children's body mass index. *American Journal of Public Health, 97*, 306-310. doi:10.2105/AJPH.2005.080499

NMEA. (2018). NMEA 2000 Edition 3.x.   Retrieved from https://www.nmea.org/content/nmea_standards/nmea_2000_ed3_10.asp

Openshaw, S. (1984). *The modifiable areal unit problem.* Norwich: Geobooks.

Pearce, Saunders, D., Allison, P., & Turner, A. (2018). Indoor and Outdoor Context-Specific Contributions to Early Adolescent Moderate to Vigorous Physical Activity as Measured by Combined Diary, Accelerometer, and GPS. *Journal of Physical Activity and Health, 15*, 40-45. doi:10.1123/jpah.2016-0638

Python Software Foundation. (2018). Python Programming Language (Version 3.4). Amsterdam. Retrieved from https://www.python.org/

R Core Team. (2018). R: A language and environment for staistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raustorp, A., Pagels, P., Boldemann, C., Cosco, N., Söderström, F., & Måtensson, F. (2012). Accelerometer measured level of physical activity indoors and outdoors during preschool time in Sweden and the United States. *J Phys Act Health, 9*.

Rissotto, A., & Tonucci, F. (2002). Freedom of Movement and Environmental Knowledge in Elementary School Children. *Journal of Environmental Psychology, 22*(1-2), 65-77. doi:10.1006/jevp.2002.0243

RStudio Team. (2015). RStudio: Integrated Development for R (Version 1.1.453). Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/.

Sallis, J. F., Cervero, R. B., Ascher, W., Henderson, K. A., Kraft, M. K., & Kerr, J. (2006). An ecological approach to creating active living communities. *Annu Rev Public Health, 27*, 297-322. doi:10.1146/annurev.publhealth.27.021405.102100

Spence, M., & White, D. (1992). *EMAP sampling grid technical report.* Retrieved from Corvallis, Oregon:

Stokols, D. (1992). Establishing and maintaining healthy environments: toward a social ecology of health promotion. *American Psychologist, 47*(1), 6-22.

Tandon, P. S., Saelens, B. E., Zhou, C., Kerr, J., & Christakis, D. A. (2013). Indoor versus outdoor time in preschoolers at child care. *Am J Prev Med, 44*. doi:10.1016/j.amepre.2012.09.052

Thierry, B., Chaix, B., & Kestens, Y. (2013). Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *Int J Health Geogr, 12*(14).

Tillmann, S., Tobin, D., Avison, W., & Gilliland, J. (2018). Mental health benefits of interactions with nature in children and teenagers: a systematic review. *J Epidemiol Community Health.* doi:10.1136/jech-2018-210436

Wang, J., He, X. G., & Xu, X. (2018). The measurement of time spent outdoors in child myopia research: a systematic review. *Int J Ophthalmol, 11*(6), 1045-1052. doi:10.18240/ijo.2018.06.24

Wu, J., Jiang, C., Houston, D., Baker, D., & Delfino, R. (2011). Automated time activity classification based on global positioning system (GPS) tracking data. *Environmental Health, 10*.

Zorzi, R., & Gagne, M. (2012). *Youth Engagement with Nature and the Outdoors: A Summary of Survey Findings.* Retrieved from https://davidsuzuki.org/wp-content/uploads/2012/09/youth-engagement-nature-outdoors.pdf

Chapter 5

5 Discussion and Conclusions

Chapters 2 to 4 comprise the substantive portion of this dissertation. The underlying theoretical basis for the substantive chapters was the ecological model of health with the built environment as an important influencer in shaping health outcomes in children. This theory, presented in Chapter 1, informed the study design and discussion for each subsequent chapter. Quantitative methods were used to measure the efficacy (and key problems) of using common address proxies, as well as to identify, classify, and bin GPS tracks to better evaluate the role that the built environment has on outdoor activity. Included in this final chapter is a summary of the contributions of the two empirical studies found in Chapters 3 and 4. Further discussion focuses on a review of the limitations of the two studies, suggestions for future research, and moving forward with policy interventions.

5.1 Summary of study findings and contributions

The contributions of the first study (Chapter 3) are twofold: first, it quantitatively describes the magnitude of positional discrepancies that result when a set of the most commonly-used home location proxies are implemented across several different neighbourhood types; and, secondly, it measures the misclassifications of the quantity of health-related facilities within local environments. The findings of this study have revealed that if commonly-used proxies such as centroids of census tracts, dissemination areas, and even postal codes, are used instead of exact addresses, positional discrepancies can be significantly large. If positional discrepancies are high, such 'ego-centric' neighbourhood units will be significantly 'off centre', and local environments can be mischaracterized and therefore lead to inaccurate assumptions. For example, the chances of misclassifying a health-promoting feature of the neighborhood, such as a recreation area, or a health-damaging feature such as a junk food outlet, as accessible (or not) can be unacceptably high, particularly when threshold distances are short, such as the commonly-used 5-min walk zone (500 m buffer). In urban neighbourhoods, when census tracts are used as home location proxies, instead of the gold-standard rooftop residential

address points, the containment misclassification of recreation areas within 500 m of the proxies is nearly half (49.5%). In rural areas the use of postal codes as an address proxy results in mean positional discrepancies to the closest recreation area of 1610 metres.

If positional discrepancies are too large, it will be impossible for the researcher to resolve whether any health effects of an environment are truly present. Even more troubling is the fact that faulty public policies may be formed and critical decisions made based on faulty environment-health research which does not take into consideration critical positional discrepancies. Improving the accuracy of the distance calculations increases the utility of the findings for making decisions and enacting policies aimed at improving a population's spatial accessibility to environmental features that contribute to their overall health and well-being. The data used is this study is ubiquitous across Canada, Spatial data is widely available including census boundary files, road networks, postal codes, lot centroids, and increasingly roof-top address points making this study relevant to all Canadian cities, small towns, and rural areas.

The key strengths of the second study (Chapter 4) emanate from the multiple strengths of the STEAM protocol, such as the large number of child participants, the use of one-second epochs on the GPS units, the high-resolution of neighbourhood built environment data, and the presence of detailed daily activity diaries. Few previous studies, except for Cooper et al. (2010) (PEACH) and McMinn et al. (2014) (PALMS) have had such a large sample of child participants. The use of the one-second epochs, in conjunction with activity diaries, time-based known locations (school day) and thematically and temporally accurate spatial reference data, in combination with the inclusion of a tessellated hexagonal surface and classification using the random forest model, adds to the feasibility and uniqueness of this study.

Studies have been conducted on the effect that seasonality plays on children's outdoor activity (Tucker & Gilliland, 2007), however, there are no known studies using GPS tracking in the Canadian context; therefore, the second study (Chapter 4) filled this gap. In addition to examining the effect of seasonality on the proportion of total time spent outdoors (engagement), the study in Chapter 4 considered other independent variables

from the *household* and *individual-levels*. Previous research has shown the utility of 'big data' - machine learning data classifiers on GPS tracks, particularly for identifying modes of travel and stops (Dwivedi & Dikshit, 2013; Ellis et al., 2014; Kestens et al., 2016; Thierry et al., 2013). Very few studies, however, have employed these types of algorithms to classify the wearer as being indoors or outdoors; exceptions are some validation studies of the PALMS application (Ellis et al., 2014; Lam et al., 2013). As mentioned in Chapter 2, the indoor vs outdoor classification part of the PALMS was limited to only the Qstarz brand of GPS receiver; however, this dissertation employed a method which resulted in higher classification match rates so that future studies where subjects are tracked with GPS receivers will not be limited to specific brand of receiver.

Following the implementation of the novel combination of methods to classify indoor vs outdoor, the study mapped when, where (exposure), and for how long (engagement) that children spent outdoors. Children in this study, on average spent 18.5% of their recorded time outdoors which translates to an average of 81 minutes (1 hour and 21) of outdoor time per day which is a similar result to a recent study of British children with a median total outdoor time of 80.3 minutes (Pearce et al., 2018).

The studies in this dissertation are interconnected through the common goal of improving methodological rigor in the measurement of children's accessibility to, exposure to, and engagement with health-related features of their environment to ultimately better our understanding of the links between environment and children's health. This dissertation therefore makes methodological and practical contributions. The methodological contribution is twofold: firstly, it informs future researchers on the best practices for utilizing address proxy level data and, secondly, it proposes a novel approach for the classification of point clouds of GPS coordinates for children's environmental health studies. The practical contribution is in the utilization of passive GPS collection in combination with ancillary spatial data to identify environments that influence children's outdoor activity levels. The outcomes of this research add to the discourse of the relationship between the built environment and children's health. It adds to the mounting evidence of the role that the design of urban and suburban environments play in the health of the people who live in them. Ultimately, this research empowers municipal

planners and policymakers with better evidence to make informed decisions regarding the planning for and design of outdoor public spaces that foster children's outdoor activity.

5.2   Synthesis of findings

5.2.1   Limitations

Despite the multiple contributions of this dissertation, there are also limitations. Firstly, the GPS devices worn by the subjects generated a series of coordinates that had varying levels of positional accuracy, both relative and absolute. These positional inaccuracies were expected and are commonly the result of the noise, bias, and blunders that persist in all studies using GPS technology. As mentioned earlier in this dissertation, positional errors occur from noise and bias when a participant wearing a GPS unit enters, exits, or remains within a blocking structure such as a building or dense tree canopy, or during the time while the unit is initially turned on. Regarding blunders, it was found that periodically the GPS units did not record for spans of time due to battery drain, and user errors.

Secondly, it is expected that spatial bias occurs while using secondary GIS data sets which are generated by agencies for their specific purposes and never meant for this study. For example, the building footprints used from the City of London data were generated for cartographic purposes rather than for engineering purposes (City of London, 2010-2013). The building polygons are product of orthographic photogrammetric digitizing methods at a large map scale, not constructed from coordinate geometry using high-resolution survey tools. What this points to is that the 'edge' of each polygon feature in the GIS building layer may not be entirely positionally accurate, nor is that positional error homogeneous. Therefore, the secondary datasets used in the creation of the hexagonal surface variables, and used in the proximity and classification methods suffer from some spatial inaccuracies. Although thematic and temporal inaccuracies are kept to a minimum in the two studies due to the use of contemporary GIS data, there will be some cases where misclassification and changes in the secondary data map features are not reflected in the data set used.

Thirdly, even though efforts are made in the methodology to reduce the modifiable area unit problem (MAUP), and uncertain geographic context problem (UGCoP), through GPS tracking and the tessellated hexagon surface, these errors still might persist in some minimal way. The two effects are expected whenever the GPS points are aggregated spatially and through that aggregation, the spatial and temporal quality of the GPS points are lost.

Lastly, this dissertation does not take into account the qualitative aspects of the ecological model, in particular, the transactional relationship between the child and their environment. For instance, the intentions and meaning a child assigns to their experiences in the physical and built environments are not addressed and are outside the scope of this dissertation.

5.2.2   Future directions

The results of this dissertation show that the quality of the spatial data used is of the upmost importance to a successful study. Therefore, it is imperative that future studies source high resolution spatial data that is thematically and temporally accurate. In Canada, researchers primarily rely on secondary data provided from government agencies at the various levels. In this dissertation, data was used from two local municipalities (City of London and County of Middlesex) in conjunction with federally published data (Statistics Canada), and data provided from a private company (DMTI Spatial Inc.). Recent advances in artificial intelligence pattern identification using remotely sensed imagery is an exciting development where researchers using these new algorithms can identify and derive a wide set of natural and built environment features with a high degree of accuracy (Lary, 2010; Zeng et al., 2013; Zeng et al., 2014). Supporting this new and exciting development is that the cost of space-borne, multi-spectral digital imagery has decreased greatly over the last decade. Perhaps even more exciting for future research applications in environment and health is the potential use of drones which carry multi-spectral scanners to provide mapping data of even higher resolution and the ability to map the landscape at the time a study is being performed. The hexagon tessellated surface does provide a more accurate measure of exposure and engagement than the most common measures which typically only measure how accessible a health

promoting/demoting feature is in the built environment, not whether that feature was seen, visited, or engaged with by a person or population of interest. This dissertation describes the methodological framework where hexagonal tessellations can be used in other urban and suburban environments in North America.

Regarding children's health studies, further data collection, both quantitative and qualitative, on the genuine actions of the children could be used to complement the findings presented in this dissertation. It would be helpful to have a form of 'member checking', using map-based interviews for example, to confirm whether or not children were actually indoors/outdoors and the reasons behind their environmental activities and decisions to be indoors/outdoors. Additionally, a natural next step in this research would be to link children's time spent outdoors, and time spent engaging in different outdoor environments, with a series of health-related behaviours, such as objective measures of vigorous physical activity, stress, or mood. Future research could compare the results using traditional methods versus the methodology forwarded in this dissertation. Finally, further research is also required to confirm our findings within different age groups, seasons, weather conditions, urban settings, and from different geographic areas within and beyond Canada.

5.3   Policy implications

There is a growing trend in public health studies, particularly within the burgeoning field of 'active living research', toward the use of 'ego-centric' units (typically defined by buffers around a study participant's residence) to characterize a participant's neighborhood and to examine the effect that local environmental factors (e.g., the mix of land uses and coverage of sidewalks) may have on health-related behaviors such as walking (Larsen et al., 2009) and outcomes such as physical activity levels (Tucker et al., 2009). The findings in Chapter 3 reveal that if commonly-used proxies such as centroids of census tracts, dissemination areas, and even postal codes, are used instead of exact addresses, positional discrepancies can be significantly large. If positional discrepancies are large, such 'ego-centric' neighbourhood units will be significantly 'off center', and local environments can be mischaracterized, leading to misclassification of 'accessibility'. For example, the chances of misclassifying a health-damaging feature

such as a junk food outlet as accessible (or not) can be unacceptably high, particularly when threshold distances are short, such as the commonly-used 500 m buffer (or 5-min walk zone). If positional discrepancies are too large, it will be impossible for a researcher to determine if any true link exists between environmental features and health-related behaviours or outcomes. The practical impact of these discrepancies not properly identified is that they could lead to important policies and/or decisions being made with poor or even erroneous evidence. Improving the accuracy of the distance calculations increases the utility of the findings for making decisions and enacting policies aimed at improving a population's spatial accessibility to features of the environment that contribute to their health and quality of life.

There is an awareness in public policy sphere that children are spending less time outdoors than ever before (Gilliland, 2018; Zorzi & Gagne, 2012) further suggested by the findings in Chapter 4, which has led to an increase in studies exploring the association between time spent outdoors and some physical and mental health outcomes. With an increase in public concerns over children's health, there is a general acceptance that the physical environment (built and natural) plays in mitigating or exacerbating health issues (Tillmann et al., 2018).

Further research is required before the results of this study are applied in any way towards policy interventions. The geographic relationships refined and studied in this thesis that correspond to the socio-ecological model's *individual, household, and neighbourhood levels* offered a clearer path to understanding the complex interactions between the built environment and children's health. The complexity of the problem, however will always be more complicated than it might seem. The problem of children spending less time outdoors is more complex than factors of seasonality, quantity of green space and varied land use types. When policy interventions occur, they should begin with a small scale intervention so that the results can be measured. The design of more appealing parks, planting more street trees, or locating health promoting facilities closer to residential areas, and the removal of health demoting facilities away from where children spend time (e.g. schools) all sound as if they might positively address part of the problem. The intervention might work, or might not work or worse be counter to what the

policy was hoped to achieve. Additional multi-disciplinary approaches in studying why children are spending less time outdoor is warranted. This thesis offers a way forward with a methodological breakthrough for overcoming the problems inherent in GPS indoor and outdoor classification, and in the linking of exposure and engagement environment variables in a meaningful way.

5.4   Conclusion

Rising public concerns over certain children's health issues, such as obesity, physical (in)activity and mental health concrens has led to a growing public and academic interest in gaining a better understanding of the role of the physical environment (built and natural) (Tillmann et al., 2018). Additionally, the public and academic realization that children are spending less time outdoors and in nature than ever before, has led to a rapid increase in studies exploring the link between time spent outdoors and/or in nature and certain physical and mental health outcomes. Nevertheless, as previous research has indicated (Tillmann et al., 2018), much of this research is difficult to decipher due to incomparable, imprecise, or inaccurate methods for assessing environmental exposure. This dissertation presents several methodological breakthroughs for overcoming the problems inherent in the literature, which should be of considerable interest to researchers and policymakers.

A quantitative socio-ecological geographic study was employed to identify data and methods to best associate children's accessibility to, exposure to, and engagement in their environment, which in turn, plays a crucial role in healthy development. This study used a spatial quantitative approach to practically measure, classify, categorize, and map children's spatial behaviours and the environmental features in their neighbourhoods. Ultimately, this dissertation offers a warning for future researchers of the folly of using some widely available spatial data for accessibility and exposure studies and also offers an improved methodology for understanding children's environmental exposures and how environmental factors might influence children's health-related behaviours and outcomes.

5.5 References

City of London. (2010-2013). *Parcels, buildings, address points, and health facilities GIS files [DVD]*.

Cooper, A. R., Page, A. S., Wheeler, B. W., Hillsdon, M., Griew, P., & Jago, R. (2010). Patterns of GPS measured time outdoors after school and objective physical activity in English children: the PEACH project. *Int J Behav Nutr Phys Act, 7*, 31. doi:10.1186/1479-5868-7-31

Dwivedi, R., & Dikshit, O. (2013). A comparison of particle swarm optimization (PSO) and genetic algorithm (GA) in second order design (SOD) of GPS networks. *Journal of Applied Geodesy, 7*(2). doi:10.1515/jag-2013-0045

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Front Public Health, 2*, 36. doi:10.3389/fpubh.2014.00036

Gilliland, J. (2018). Lawson Foundation Systematic Review.

Kestens, Y., Thierry, B., & Chaix, B. (2016). Re-creating daily mobility histories for health research from raw GPS tracks: Validation of a kernel-based algorithm using real-life data. *Health Place, 40*, 29-33. doi:10.1016/j.healthplace.2016.04.004

Lam, M., Godbole, S., Chen, J., Oliver, M., Badland, H., Marshall, S. J., . . . Kerr, J. (2013). *Measuring Time Spent Outdoors Using a Wearable Camera and GPS.* Paper presented at the SenseCam '13 Proceedings of the 4th International SenseCam & Pervasive Imaging Conference, New York.

Larsen, K., Gilliland, J., Hess, P., Tucker, P., Irwin, J., & He, M. (2009). The influence of the physical environment and sociodemographic characteristics on children's mode of travel to and from school. . *American Journal of Public Health, 99*(3), 520-526. doi:10.2105/AJPH.2008

Lary, D. (2010). Artificial intelligence in geoscience and remote sensing. *Artificial Intelligence in Geoscience and Remote Sensing, Geoscience and Remote Sensing*.

McMinn, D., Oreskovic, N. M., Aitkenhead, M. J., Johnston, D. W., Murtagh, S., & Rowe, D. A. (2014). The physical environment and health-enhancing activity during the school commute: Global positioning system, geographical information systems and accelerometry. *Geospatial Health, 8*, 569-572. doi:10.4081/gh.2014.46

Pearce, Saunders, D., Allison, P., & Turner, A. (2018). Indoor and Outdoor Context-Specific Contributions to Early Adolescent Moderate to Vigorous Physical Activity as Measured by Combined Diary, Accelerometer, and GPS. *Journal of Physical Activity and Health, 15*, 40-45. doi:10.1123/jpah.2016-0638

Thierry, B., Chaix, B., & Kestens, Y. (2013). Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *Int J Health Geogr, 12*(14).

Tucker, P., & Gilliland, J. (2007). The effect of season and weather on physical activity: a systematic review. *Public Health, 121*(12), 909-922. doi:10.1016/j.puhe.2007.04.009

Tucker, P., Irwin, J. D., Gilliland, J., He, M., Larsen, K., & Hess, P. (2009). Environmental influences on physical activity levels in youth. *Health Place, 15*(1), 357-363. doi:10.1016/j.healthplace.2008.07.001

Zeng, C., Wang, J., & Lehrbass, B. (2013). An Evaluation System for Building Footprint Extraction From Remotely Sensed Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(3), 1640-1652. doi:10.1109/jstars.2013.2256882

Zeng, C., Wang, J., Zhan, W., Shi, P., & Gambles, A. (2014). An elevation difference model for building height extraction from stereo-image-derived DSMs. *International Journal of Remote Sensing, 35*(22), 7614-7630. doi:10.1080/01431161.2014.975375

Zorzi, R., & Gagne, M. (2012). *Youth Engagement with Nature and the Outdoors: A Summary of Survey Findings*. Retrieved from https://davidsuzuki.org/wp-content/uploads/2012/09/youth-engagement-nature-outdoors.pdf

Appendices

Appendix A: Research Ethics Approval Form - Human Participants STEAM I

Office of Research Ethics

The University of Western Ontario
Room 4180 Support Services Building, London, ON, Canada N6A 5C1

Western

Use of Human Subjects - Ethics Approval Notice

Principal Investigator: Dr. J. Gilliland

Review Number: 16495   Review Level: Full Board

Review Date: August 11, 2009

Protocol Title: Emerging Methodologies for Examining Environmental Influences on Children's Exposure to Air Pollution

Department and Institution: Geography, University of Western Ontario

Sponsor:

Ethics Approval Date: December 01, 2009   Expiry Date: August 31, 2010

Documents Reviewed and Approved: UWO Protocol, Letter of Information and Consent, Assent.

Documents Received for Information:

This is to notify you that The University of Western Ontario Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement: Ethical Conduct of Records Involving Humans and the applicable laws and regulations of Ontario has granted approval to the above named research study on the approval date noted above.

This approval shall remain valid until the expiry date noted above assuming timely and acceptable responses to the NMREB's periodic requests for surveillance and monitoring information. If you require an updated approval notice prior to that time you must request it using the UWO Updated Approval Request Form.

During the course of the research, no deviations from, or changes to, the study or consent form may be initiated without prior written approval from the NMREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of monitor, telephone number). Expedited review of minor change(s) in ongoing studies will be considered. Subjects must receive a copy of the signed information/consent documentation.

Investigators must promptly also report to the NMREB:

a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to this office for approval.

Members of the NMREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the NMREB.

Chair of NMREB: Dr. Jerry Paquette

This is an official document. Please retain the original in your files.   cc: ORE File

Appendix B: Research Ethics Approval Form - Human Participants STEAM II

Use of Human Participants - Ethics Approval Notice

Research

Western

Principal Investigator: Dr. Jason Gilliland
Review Number: 17918S
Review Level: Delegated
Approved Local Adult Participants: 1200
Approved Local Minor Participants: 1200
Protocol Title: Identifying casual effects on the built environment on physical activity, diet, and obesity among children.
Department & Institution: Social Science\Geography, University of Western Ontario
Sponsor: Canadian Institutes of Health Research
Heart and Stroke Foundation of Canada

Ethics Approval Date: June 08, 2011   **Expiry Date:** August 31, 2014

Documents Reviewed & Approved & Documents Received for Information:

Document Name	Comments	Version Date
Other	Revised Healthy Neighbourhood Survey for Parents.	
Other	Revised Health Neighbourhoods Survey for Youth	
Other	Revised Activity and Travel Diary for School Days and Weekend Days.	

This is to notify you that The University of Western Ontario Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans and the applicable laws and regulations of Ontario has granted approval to the above referenced revision(s) or amendment(s) on the approval date noted above.

This approval shall remain valid until the expiry date noted above assuming timely and acceptable responses to the NMREB's periodic requests for surveillance and monitoring information.

████████████ who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussions related to, nor vote on, such studies when they are presented to the NMREB.

The Chair of the NMREB is Dr. Riley Hinson. The NMREB is registered with the U.S. Department of ████████ under the IRB registration ████████

Ethics Officer to Contact for Further Information

✓ Grace Kelly	___ Janice Sutherland

This is an official document. Please retain the original in your files.

Appendix C: Copyright Release from Publication

Chapter 3:
Healy, M. A., & Gilliland, J. A. (2012). Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research. Spat Spatiotemporal Epidemiol, 3(1), 55-67. doi:10.1016/j.sste.2012.02.006

Glenn, Louise (ELS)
11:59 AM (10 minutes ago)
to xxx

Dear Mr Healy,

Thanks for your query. This is completely fine, we would just ask that you also include a link to the final published paper on ScienceDirect in all publication formats of the thesis.

https://doi.org/10.1016/j.sste.2012.02.006

All the best,

Louise

-----Original Message-----
From: xxxx
Sent: 21 August 2018 23:37
To: xxxx
Subject: Enquiry: Permission to include published article in my Dissertation

  I request your permission so that I may include my paper titled "Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research." published in Spatial and Spatio-temporal Epidemiology 3(1) 55-67 in my PhD thesis.

Best regards,
Martin Healy
PhD Candidate
University of Western Ontario
London, Ont.
Canada

Appendix D: Pre-Processing Scripts for Random Forest

```
###=============================
### Program: Get Buildings using SphereLab Stops
### Author: Martin Healy
### Date: March 18, 2018
### Description: This program processes the output points of the
### Thierry et al. (2013) Activity place detection algorithm
### for GPS data :
###      1. Codes all stops inside study area (city boundary)
###         and uses only stops with a duration > 2 min
###      2. Selects the buildings close to each valid stop
###      3. Identifies and stores the valid buildings by Participant ID
###
### Dependencies: arcpy, sys, os, SteamModules
### -------------------------------------
### 1. Successful processing of Routes and Stops from GPS Data SphereLab
###    Activity place detection algorithm for GPS data (Thierry et al.,2013)
###    using the custom HEAL Lab STEAM ArcToolBox Toolset.
### 2. Programmed to run from an ArcToolBox Script Tool
###=============================

import arcpy, sys, os, SteamModules

# Read in Variables from ArcGIS Pro
InStopBldgs = arcpy.GetParameterAsText(0)
InCity = arcpy.GetParameterAsText(1)
InStops = arcpy.GetParameterAsText(2)
OutputBldgwithStops = arcpy.GetParameterAsText(3)
arcpy.env.workspace = arcpy.GetParameterAsText(4)

arcpy.env.overwriteOutput = True

#Local Variables

CityStops = "in_memory\\InCityStops"
CityStopsGT2min = "in_memory\\InCityStopsGT2min"
NearStopsTab= "in_memory\\NearStopsTable"

# Start Processing
arcpy.AddMessage("Processing CLIP in City and Stops > 2 mins")
# Create the output Buildings for stops empty Feature Class and add fields
arcpy.CreateFeatureclass_management(SteamModules.GetPath(OutputBldgwithStops),
SteamModules.GetFCNamefromPath(OutputBldgwithStops),"POLYGON",InStopBldgs)
arcpy.AddFields_management(OutputBldgwithStops,[["SID", "TEXT", "", 10],
                                    ["DayType", "TEXT", "", 10],
                                    ["STOPID", "TEXT", "", 20],
                                    ["BLDGID", "LONG"]])
# clip only stops within London and keep only the stops with duration over 2
mins
arcpy.Clip_analysis(InStops, InCity, CityStops)
arcpy.Select_analysis(CityStops, CityStopsGT2min, 'NbTicks > 120')

arcpy.AddMessage("Processing NEAR Stops to Buildings")
```

```python
# find the closest building to each stop
arcpy.GenerateNearTable_analysis(CityStopsGT2min,InStopBldgs,NearStopsTab,"","
","","CLOSEST")
arcpy.JoinField_management(CityStopsGT2min,"OBJECTID",NearStopsTab,"IN_FID")
arcpy.AddField_management(CityStopsGT2min, "BLDGID", "LONG")
arcpy.CalculateField_management(CityStopsGT2min, "BLDGID","!NEAR_FID!" )

arcpy.AddMessage("Looping through buildings and Stops to Generate " +
OutputBldgwithStops +" File")
StopsCursor = arcpy.da.SearchCursor(CityStopsGT2min,["SID", "STOPID",
"DayType","BLDGID","NEAR_DIST"])
BuildingCursor = arcpy.da.SearchCursor(InStopBldgs,["SHAPE@","OBJECTID"])
InsCursor = arcpy.da.InsertCursor(OutputBldgwithStops,
        ["SHAPE@","SID","DayType","STOPID","BLDGID"])

# Use the stops with building OBJECTID on outside loop and buildings on inside
# when the NEAR_FID = Building ObjID then write the
arcpy.MakeTableView_management(CityStopsGT2min, "myTableView")
count = int(arcpy.GetCount_management("myTableView").getOutput(0))
cnt = 0
for rec in StopsCursor:

    for blg in BuildingCursor:
        if rec[3] == blg[1]:
            InsCursor.insertRow((blg[0],rec[0],rec[2],rec[1],rec[3]))
            break
    BuildingCursor.reset()
    cnt += 1

    arcpy.AddMessage(str(float(cnt/count) * 100) + "% complete")

del InsCursor
del BuildingCursor
del StopsCursor
```

```
###=============================
### Program:  STEAM GPS Pre-Processer for R
### Author: Martin Healy
### Date: April 11, 2018
### Description: This program formats and prepares the STEAM GPS data stored
### in a File Geodatabase for processing random forest in R: The program does
### the following:
###        1. Flags only GPS points inside study area (within city boundary)
###        2. Creates a custom variable (DistFromBuild) by measures distance
###           from all GPS points to blocking structures (buildings) and
###           isolated the only the School Building(s) distance to be used
###           during validated school times
###        3. Sets Training Flag and Indoor/Outdoor classes, but validates them
###           with tests for Indoor recesses (from diary), or for times the
###           child is off school grounds during scheduled school times.
###        4. Identifies SphereLab Routes as active and vehicular travel. The
###           GPS points used to create the routes are flagged as training data
###           and coded as Indoors (vehicular travel), or Outdoors (active
###           travel).
###        5. Creates a custom variable (DisttoLastPoint) by measuring the
###           distance between successive points.
### Can be run standalone or using an ArcTool Script Tool
### Dependencies: arcpy, sys, os, gc, SteamModules
###-------------------------------
### Only to run following the:
### 1. Successful processing of Routes and Stops from GPS Data SphereLab
###    Activity place detection algorithm for GPS data (Thierry et al.,2013)
###    using the custom HEAL Lab STEAM ArcToolBox Toolset.
### 2. Identification/creation of the school building polygons and successful
###    processing of the building identification/coding from the Spherelab
###    stops using script Get Buildings using SphereLab Stops accessed from
###    the HEAL Lab STEAM ArcToolBox Toolset.
###=============================
import arcpy, sys, os, gc
import SteamModules

arcpy.env.overwriteOutput = True

arcpy.env.workspace = arcpy.GetParameterAsText(0)
InStopBldgs = arcpy.GetParameterAsText(1)
InSchBldgs = arcpy.GetParameterAsText(2)
InBoundary = arcpy.GetParameterAsText(3)
InDiaryTab = arcpy.GetParameterAsText(4)
InActiveRoutesFC = arcpy.GetParameterAsText(5)
InVehicleRoutesFC = arcpy.GetParameterAsText(6)
OutCSVFile = arcpy.GetParameterAsText(7)
Season = arcpy.GetParameterAsText(8)
outputWorkspace = arcpy.GetParameterAsText(9)
OutputStatsFolder = arcpy.GetParameterAsText(10)
outDataPathTable = arcpy.GetParameterAsText(12)

## Set Environment Variables
ws = arcpy.env.workspace
Scratch_workspace = "in_memory\\"
test_workspace = "C:\\w\PhDSTEAMProcessing\\RProcessing\\Scratchy.gdb"
```

```python
try:

# Get the list of Feature Datasets from a Workspace
    datasets = arcpy.ListDatasets(feature_type="Feature")
    datasets = [''] + datasets if datasets is not None else []

# Loop through a list of feature classes in the workspace
    FC_Count = 1
    for ds in datasets:
        for fc in arcpy.ListFeatureClasses("*", "Point", ds):
            print("Processing: " + fc)
            InputGPSpath = os.path.join(arcpy.env.workspace, ds, fc)

        #  Set the output name VARIABLE to be the same as the input name, and
        #   locate in the 'temp_workspace' workspace
            OutputGPSFCpath = os.path.join(outputWorkspace, fc)
            outputFC_mem = os.path.join("in_memory", fc)

        #  Get SID value
            SidID = fc.upper()
            SchoolID = SteamModules.GetSchoolID(SidID,Season)
            inSchBldg = "in_memory\\SchBldg"
            anExpression = "SchoolID = '" + SchoolID + "'"
            arcpy.Select_analysis(InSchBldgs, inSchBldg, anExpression )
            NumRecords = int(arcpy.GetCount_management
                    (inSchBldg).getOutput(0))
            if NumRecords == 0:
                print("No School Buildings were selected for "
                    + SchoolID + " INVESTIGATE")
                break
            GPSLayer = "lyr_" + str(SidID)

            # Make a Copy of the FC and save in_memory.  Add the four fields
             # (Indoors,DistFromBuild,INDOORS,TrainingFlag)
            outputFC_InCity = SteamModules.SetInCityFlag(fc,InBoundary,SidID,
                    GPSLayer,Season,outputFC_mem)

            ### Get all the nearest buildings identified from SphereLab Stops
            ### for all GPS points
            outputFC_InCity_AllBldgs = SteamModules.GetNearestBuilding
                    (outputFC_InCity,InStopBldgs,SidID,"ALL")
            if outputFC_InCity_AllBldgs == "Error"
                    or outputFC_InCity_AllBldgs == False:
                break
            ### Set TrainingFlag for all scheduled times
            outputFC_Training = SteamModules.SetTrainingFlag
                    (outputFC_InCity_AllBldgs,SidID)
            if outputFC_Training == False:
                break
            ### Add Field to hold if the GPS tracks are Travelling during
            ### school time, if so the distance to school building will not
            ### override the general distance to buildings
            arcpy.management.AddField(outputFC_Training,"offcampus","SHORT")
```

```python
            ### Refine the Training Data TrainingFlag remove when children
             ### off-campus
            inTrainStatsTable = "in_memory\\TrainStatsTable"
            outputFC_Training_no_OC = SteamModules.RemoveOffCampus
                    (outputFC_Training, inTrainStatsTable, Season)
            if outputFC_Training_no_OC == False:
                break
            ### Refine the Training Data toggle INSIDE, if children was Indoor
             ### for Recess
            outputFC_Training_OKRecess = SteamModules.RefineIndoorRecess
                    (outputFC_Training_no_OC ,SidID,InDiaryTab ,Season)
            if outputFC_Training_OKRecess == False:
                break

            ### Refine the Training Data TrainingFlag for Active-non-active
            ### travel Routes to, remove children off-campus, not
            ### outside for recess
            outputFC_Training_travel = SteamModules.RefineTravel
                    (outputFC_Training_OKRecess, InActiveRoutesFC,
                    SidID, Season, "ACTIVE")
            if outputFC_Training_travel == False:
                break
            outputFC_Training_travelall = SteamModules.RefineTravel
                    (outputFC_Training_travel, InVehicleRoutesFC,
                    SidID, Season, "VEHICLE")
            if outputFC_Training_travelall == False:
                break
            ### Get the school Building(s) as nearest buildings for all in
             ### school Training GPS points
            outputFC_Training_mem = SteamModules.GetNearestBuilding
                    (outputFC_Training_travelall,inSchBldg,
                    SidID,"TRAINING")
            if outputFC_Training_mem == False:
                break

            ### Add Distance to previous point
            outputFC_withDistFromLast = SteamModules.CalcDistToLastPoint
                    (outputFC_Training_mem)
            if outputFC_withDistFromLast == False:
                break
# Make a copy of the GPS Points ready for use in R
            arcpy.Select_analysis(outputFC_Training_mem, OutputGPSFCpath)

            SteamModules.CreatePathTableForR(outDataPathTable,
                    OutputGPSFCpath,SidID, Season,outputWorkspace)

            FC_Count = FC_Count + 1
# Clear variable memory
            inSchBldg = None
            GPSLayer = None
            outputFC_InCity = None
            outputFC_InCity_AllBldgs = None
            outputFC_Training_no_OC = None
            outputFC_Training_travel = None
            outputFC_Training_OKRecess = None
```

```python
            outputFC_Training_mem = None
            inTrainStatsTable = None
            gc.collect()
except (RuntimeError, TypeError, NameError, IOError) as err:
    print("Oops! Error in the PrepSTEAMGPSDataForR module: " + err)
```

```python
###============================
### Program: SteamModules
### Author: Martin Healy
### Date: March 29, 2018
### Description: This program contains the modules for all STEAM Python
###             Scripts
###
### Dependencies: arcpy, os, csv
###============================

import arcpy, os, csv, math

# Get Number of GPS records
def GetFeatureCount(inSID, inFCLayer):
    NumRecords = int(arcpy.GetCount_management(inFCLayer.getOutput(0)))
    if NumRecords == 0:
        print("No GPS Points were selected for " + inSID + " INVESTIGATE")
    return ""

# Get School ID from name of participant ID (SID)
def GetSchoolID(inSID, Season):
    if len(inSID) == 5:
        if Season == "S2010" or Season == "W2010":
            outschool = inSID[0] + inSID[1]
        else:
            outschool = inSID[0]
    elif len(inSID) == 6:
        outschool = inSID[0] + inSID[1]
    else:
        outschool = "Error in name of School From SID"
    return outschool

# Some GPS Feature Class tables do not include standard names
# make them standard.
def HasDayTypeField(inFC):
    try:
        boolVar = False
        fList = arcpy.ListFields(inFC)
        for f in fList:
            if f.name == "DayType" or f.name == "DAYTYPE":
                boolVar = True
        return boolVar
    except (RuntimeError, TypeError, NameError, IOError) as err:
        print("Oops! Error in the HasDayTypeField module: " + err)

# Flag only GPS records within a boundary
def SetInCityFlag(fc, inCityBnd, SIDID,GPSLayer,Season,outputFC):
    try:
        arcpy.Select_analysis(fc, outputFC)
        if HasDayTypeField(outputFC):
            arcpy.management.AddFields(outputFC, [["DistFromBuild",
                "DOUBLE"],["DistFromLastPoint","DOUBLE"],
                ["TrainingFlag","SHORT"],["INDOOR","SHORT"],
                ["InCity","SHORT"],["INOUT","SHORT"]])
        else:
```

```python
            if Season == "S2010" or Season == "W2010":
                arcpy.management.AddFields(outputFC,  [
                    ["DayType","TEXT","",10],["DistFromBuild", "DOUBLE"],
                    ["DistFromLastPoint","DOUBLE"],["TrainingFlag","SHORT"],
                    ["INDOOR","SHORT"],["InCity","SHORT"],["INOUT","SHORT"]])
                if Season == "W2010":

                    arcpy.CalculateField_management(outputFC,"DayType",'!DAY!'
                        , "PYTHON3")
                else:

                    arcpy.CalculateField_management(outputFC,"DayType",'!Day_T
                        ype!', "PYTHON3")
            else:
                print("No DAYTYPE Field in GPS Data: " + err)
        arcpy.CalculateField_management(outputFC,"TrainingFlag",0, "PYTHON3")
        arcpy.CalculateField_management(outputFC,"InCity",0, "PYTHON3")
        arcpy.MakeFeatureLayer_management(outputFC,GPSLayer)
        arcpy.SelectLayerByLocation_management(GPSLayer,
            "INTERSECT",inCityBnd)
        arcpy.CalculateField_management(GPSLayer,"InCity",1, "PYTHON3")
        arcpy.SelectLayerByAttribute_management(GPSLayer,"CLEAR_SELECTION")

        return outputFC
    except (RuntimeError, TypeError, NameError, IOError) as err:
        print("Oops! Error in the SetInCityFlag module: " + err)

# Get the distance to the closest buildings identified as stops
def GetNearestBuilding(inGPSFC, inBuildings, SidID, Type):
    try:
        if Type == "ALL":
            allSIDBlg = "in_memory\\allSIDBlg"
            #Calc dist to all buildings and set the Training Flag = 0
            anExpression = "SID = '()'".format(SidID)
            arcpy.Select_analysis(inBuildings,allSIDBlg, anExpression)
            NumRecords =int(arcpy.GetCount_management(allSIDBlg).getOutput(0))
            if NumRecords == 0:
                print("No Buildings were selected for " + SidID + "
                    INVESTIGATE")
                return "Error"

            arcpy.Near_analysis(inGPSFC,allSIDBlg)
            arcpy.CalculateField_management(inGPSFC,"DistFromBuild","!NEAR_DI
                ST!", "PYTHON3")
            arcpy.DeleteField_management(inGPSFC,["NEAR_DIST", "NEAR_FID"])
            arcpy.CalculateField_management(inGPSFC, "INDOOR", 0, "PYTHON3")
            arcpy.CalculateField_management(inGPSFC,"TrainingFlag", 0,
             "PYTHON3")

        elif Type == "TRAINING":

        #Calc dist to school building(s)

            inGPSLayer = "lyrschool_inGPSFC"
            arcpy.MakeFeatureLayer_management(inGPSFC,inGPSLayer)
```

```python
            arcpy.Near_analysis(inGPSLayer,inBuildings)
            Trainingwclause1 = "TrainingFlag = 1"
            Trainingwclause2 = "TIME_BLOCK_NAME = 'AM_School1' OR
                            TIME_BLOCK_NAME = 'PM_School1' OR TIME_BLOCK_NAME =
                            'AM_School2' OR TIME_BLOCK_NAME = 'PM_School2'"
            Trainingwclause3 = "TIME_BLOCK_NAME = 'AM_Recess' OR
                            TIME_BLOCK_NAME = 'PM_Recess'"

            # Not on campus for the day or part of day (travelling)
            Trainingwclause4 = "offcampus = 1"
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "NEW_SELECTION",Trainingwclause1)
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "ADD_TO_SELECTION", Trainingwclause2)
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "ADD_TO_SELECTION",Trainingwclause3)
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "REMOVE_FROM_SELECTION",Trainingwclause4)

        # if the DistFromBuild is much less than the distance to school
        # (<800m), then keep old DistFromBuild value. The child is off
        # campus, not shown as a route, but close to school

            Trainingwclause5 = "DistFromBuild < (NEAR_DIST - 800)"
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "REMOVE_FROM_SELECTION",Trainingwclause5)
            arcpy.CalculateField_management(inGPSLayer,
                    "DistFromBuild","!NEAR_DIST!", "PYTHON3")
            arcpy.DeleteField_management(inGPSLayer,["NEAR_DIST", "NEAR_FID"])
            arcpy.SelectLayerByAttribute_management(inGPSLayer,
                    "CLEAR_SELECTION")
        return inGPSFC
    except (RuntimeError, TypeError, NameError, IOError) as err:
        print("Oops! Error in the GetNearestBuilding module: " + err)

# Set the Training Flag for R Random Forest Creation
def SetTrainingFlag(inGPSFC, SidID):
    try:
        inGPSLayer = "lyr_inGPSFC"
        arcpy.MakeFeatureLayer_management(inGPSFC,inGPSLayer)
        Trainingwclause1 = "TIME_BLOCK_NAME = 'AM_School1'
            OR TIME_BLOCK_NAME = 'PM_School1' OR
            TIME_BLOCK_NAME = 'AM_School2' OR TIME_BLOCK_NAME = 'PM_School2'"
        Trainingwclause2 = "TIME_BLOCK_NAME = 'AM_Recess'
            OR TIME_BLOCK_NAME = 'PM_Recess'"
        # set the Training Flag = 1 when regimented school times
        arcpy.SelectLayerByAttribute_management(inGPSLayer,
            "NEW_SELECTION",Trainingwclause1)
        arcpy.CalculateField_management(inGPSLayer, "INDOOR", 1)
        arcpy.CalculateField_management(inGPSLayer,
            "TrainingFlag", 1, "PYTHON3")
        arcpy.SelectLayerByAttribute_management(inGPSLayer,
            "NEW_SELECTION",Trainingwclause2)
        arcpy.CalculateField_management(inGPSLayer, "INDOOR", 0, "PYTHON3")
```

```python
            arcpy.CalculateField_management(inGPSLayer,
                "TrainingFlag", 1, "PYTHON3")
            arcpy.SelectLayerByAttribute_management(inGPSLayer,"CLEAR_SELECTION")

            return inGPSFC
        except (RuntimeError, TypeError, NameError) as err:
            print("Oops! Error in the RemoveIndoorRecess Function: " + err)
            return False

# Remove any non-conforming GPS Training points (i.e. stay indoors at Recess)
def RefineIndoorRecess(inGPSFC,SID, DiaryTable, Season):
    try:
        # Choose the Diary records for the participant and Season
        sqlexpress = "SID = '()' AND S = '()'".format(SID,Season)

        #Get the Diary records for this participant for this season
        with arcpy.da.SearchCursor(DiaryTable,["SID","DayType",
            "TimeBlock","TB_IO1"],sqlexpress ) as SCursor:
            DictList = []
            for row in SCursor:
                diaryDaytype = row[1]
                diaryTimeBlock = row[2]
                diaryRecessInout = row[3]
                recessDayDict = dict(daytype = diaryDaytype,
                    timeB = diaryTimeBlock,RecessInout = diaryRecessInout)
                DictList.append(recessDayDict)

        #Do for only all the indoor recesses -> set INDOOR = 1...keep as
        # training data
        cnt = 0
        gpscnt = 0
        for aDict in DictList:

            if aDict['RecessInout'] == 2:

                if Season == "S2010" or Season == "W2010":
                    sqlexpress2 = "DayType = '()' AND
                        TIME_BLOCK_NAME = '()'"
                        .format(aDict['daytype'],aDict['timeB'])
                else:
                    sqlexpress2 = "DAYTYPE = '()' AND
                        TIME_BLOCK_NAME = '()'"
                        .format(aDict['daytype'],aDict['timeB'])

                UCursor = arcpy.da.UpdateCursor(inGPSFC,["INDOOR"],sqlexpress2
)

                for uRow in UCursor:
                    uRow[0] = 1
                    UCursor.updateRow(uRow)
                    gpscnt += 1
                cnt += 1
        print("() in () had () recess inside for a total
            of () GPS points".format(SID,Season,cnt,gpscnt))
        return inGPSFC
```

```python
        except (RuntimeError, TypeError, NameError) as err:
            print("Oops! Error in the RefineIndoorRecess Function: " + err)
            return False

# Remove any non-conforming GPS Training points (i.e. leave school Property
#during school day, stays at home)
def RemoveOffCampus(inGPSFC,TrainStatsTemp,Season):
    try:
        # Select the Training Data
        inTrainGPSLayer = "lyrtrain_inGPSFC"
        BadDayList = []
        ValidDayList =["WD1","WD2","WD3","WD4","WD5","WD6","WD7"]
        if Season == "S2010" or Season == "W2010":
            arcpy.Statistics_analysis(inGPSFC,
                    TrainStatsTemp,[["DistFromBuild","MEAN"],
                    ["DistFromBuild","RANGE"]],["DayType","TrainingFlag"])
        else:
            arcpy.Statistics_analysis(inGPSFC,
                    TrainStatsTemp,[["DistFromBuild","MEAN"],
                    ["DistFromBuild","RANGE"]],["DAYTYPE","TrainingFlag"])
        TrainTabCursor = arcpy.da.SearchCursor(TrainStatsTemp,["*"])

        # By Day, find the mean distance, if > 500m, then student is off
        # campus for part of the day, and that day will not be included in the
        # training #dataset - set TrainingFlag = 0

        # Create a list of daytypes for this participant where they are not on
        # school property during school days
        for TRec in TrainTabCursor:
            if TRec[2] == 1:
                for daytype in ValidDayList:
                    if TRec[4] > 500 and TRec[1] == daytype or
                            TRec[5] > 500 and TRec[1] == daytype:
                        BadDayList.append(TRec[1])

        # select and calculate TrainingFlag = 0 for that DayType
        if len(BadDayList) > 0:
            arcpy.MakeFeatureLayer_management(inGPSFC,inTrainGPSLayer)
            for bDay in BadDayList:
                if Season == "S2010" or Season == "W2010":
                    arcpy.SelectLayerByAttribute_management(inTrainGPSLayer,
                            "NEW_SELECTION","DayType = '" + bDay + "'")
                else:
                    arcpy.SelectLayerByAttribute_management(inTrainGPSLayer,
                            "NEW_SELECTION","DAYTYPE = '" + bDay + "'")
                arcpy.CalculateField_management(inTrainGPSLayer,
                        "TrainingFlag",0, "PYTHON3")
                arcpy.CalculateField_management(inTrainGPSLayer,
                        "INDOOR", 0, "PYTHON3")
                arcpy.CalculateField_management(inTrainGPSLayer,
                        "offcampus", 1, "PYTHON3")
                arcpy.SelectLayerByAttribute_management(inTrainGPSLayer,
                        "CLEAR_SELECTION")
```

```python
            return inGPSFC

    except (RuntimeError, TypeError, NameError) as err:
        print("Oops! Error in the RemoveOffCampus Function: " + err)
        return False

# Set Vehicle travel as INDOOR and non-vehicle as OUTDOOR
def RefineTravel(inGPSFC, inRouteFC, SID, Season,Type):
    try:
        # Use every Route - the FC has been preprocessed to remove
        # erroneous routes
        sqlexpress = "SID = '()'".format(SID)
        with arcpy.da.SearchCursor(inRouteFC,["SID",
            "DayType", "UTCStartTime_Date",
            "UTCStopTime_Date"], sqlexpress) as SCursor:
            DictList = []
            for row in SCursor:
                routeDaytype = row[1]
                routeStartTime = row[2]
                routeStopTime = row[3]
                routeDayDict = dict(daytype = routeDaytype,
                    starttime = routeStartTime, stoptime = routeStopTime)
                DictList.append(routeDayDict)


        # Do for only all the indoor recesses -> set INDOOR = 1 and set
        # training flag
        for aDict in DictList:

            if Season == "S2010" or Season == "W2010":
                sqlexpressa = "DayType = '" + aDict['daytype'] + "'"
                sqlexpressb = sqlexpressa + " AND FULLTIME >= timestamp " +
                    "'" + aDict['starttime'].strftime('%Y-%m-%d %H:%M:%S')
                     + "'"
                sqlexpress2 = sqlexpressb + " AND FULLTIME <= timestamp " +
                    "'" + aDict['stoptime'].strftime('%Y-%m-%d %H:%M:%S')
                        + "'"
            else:
                sqlexpressa = "DAYTYPE = '" + aDict['daytype'] + "'"
                sqlexpressb = sqlexpressa + " AND FULLTIME >= timestamp " +
                        "'" + aDict['starttime']
                        .strftime('%Y-%m-%d %H:%M:%S') + "'"
                sqlexpress2 = sqlexpressb + " AND FULLTIME <= timestamp " +
                        "'" + aDict['stoptime']
                        .strftime('%Y-%m-%d %H:%M:%S') + "'"

            UCursor = arcpy.da.UpdateCursor(inGPSFC,
                    ["SPEED","INDOOR","TrainingFlag","offcampus"],sqlexpress2)

            # Set as Outdoor and TrainingFlag = 1
            if Type == "ACTIVE":

                # Check if route misidentified as Active i.e. when 4 or more
                # GPS points making the route contain excessive speeds
                overspeedcnt = 0
```

```python
                for uRow in UCursor:
                    if uRow[0] > 18:
                        overspeedcnt += 1
                UCursor.reset()
                # Set Active Travel as Outdoors and TrainingFlag = 1
                if overspeedcnt < 4:
                    for uRow in UCursor:
                        uRow[1] = 0
                        uRow[2] = 1
                        uRow[3] = 1
                        UCursor.updateRow(uRow)

            # Set as Indoor and TrainingFlag = 1
            elif Type == "VEHICLE":

                for uRow in UCursor:
                    uRow[1] = 1
                    uRow[2] = 1
                    uRow[3] = 1
                    UCursor.updateRow(uRow)

        return inGPSFC
    except (RuntimeError, TypeError, NameError) as err:
        print("Oops! Error in the RefineTravel Function: " + err)
        return False

# Create the <STEAMSeason>_FCPaths Table for R
def CreatePathTableForR(tablePath,FcPath,SID,Season,outputWorkspace):

    TableCur = arcpy.da.InsertCursor(tablePath,
        ["SID","Season","dPath","ErrorCSVPath","ForestSizeInfluencePlotPath",
        "VariableImportancePlotPath","ConfusionMatrixCSVPath","tPath"])
    aPath = GetPathtoFolder(FcPath)
    ErrorCSVPath = aPath + "\\RandomForestErrors\\" + Season + "_RFError.csv"
    ForestSizePlotsPath = aPath +
        "\\OutputPlots\\ForestSizeInfluencePlotPath\\"
        + Season + "\\" + SID + "_FSIP.pdf"
    ImportancePlotsPath = aPath +
        "\\OutputPlots\\VariableImportancePlotPath\\" +
        Season + "\\" + SID + "_VIP.pdf"
    ConfuseMatrixPath = aPath +
        "\\OutputPlots\\ConfusionMatrixCSVPath\\" +
        Season + "\\" + SID + "_CM.csv"
    outTable = outputWorkspace + "\\" + SID + "inout"

    TableCur.insertRow([SID,Season,FcPath,ErrorCSVPath,
        ForestSizePlotsPath,ImportancePlotsPath,ConfuseMatrixPath,outTable])

    del TableCur

# Get the Feature Class Name from a full path to the file
def GetFCNamefromPath(inPath):
    listPath = inPath.split(os.sep)
    strFCName = listPath[len(listPath) - 1]
    return strFCName
```

```python
# Get the File path to a dataset
def GetPath(inPath):
    listPath = inPath.split(os.sep)
    for index in range(0,len(listPath)-1):
        if index == 0:
            strOutputPath = listPath[index]
        else:
            strOutputPath = strOutputPath + os.sep + listPath[index]
    return strOutputPath

# Get the File path to a workspace folder
def GetPathtoFolder(inPath):
    listPath = inPath.split(os.sep)
    for index in range(0,len(listPath)-1):
        if ".gdb" in listPath[index]:
            break
        elif index == 0:
            strDirPath = listPath[index]
        else:
            strDirPath = strDirPath + os.sep + listPath[index]
    return strDirPath

# Generates distance between two points - pass in 2 successive points
def PointDist(pt1, pt2):
    # Input: two tuples (x,y) that defined a pair of successive points
    # sqrt((x1-x2)**2 + (y1-y2)**2
    if ((pt2[0] - pt1[0]) != 0) and ((pt2[1] - pt1[1]) != 0):
        ptDist = math.sqrt(math.pow(pt2[0] - pt1[0], 2) + math.pow(pt2[1] -
pt1[1], 2))
    else:
        ptDist = 0
    return ptDist

# Calculate distance between two GPS points
def CalcDistToLastPoint(inGPSFC):
    try:
        # Create a list of coordinate tuples from Search Cursor
        pointList =[]


        # Get the points
        with arcpy.da.SearchCursor(inGPSFC,["SHAPE@XY"]) as SCursor:
            for row in SCursor:
                ptcoord = (row[0][0],row[0][1])
                pointList.append(ptcoord)

        ptCount = len(pointList)
        cnt = 0
        pDistList = []
        for firstpt in pointList:
            if  cnt == 0:
                pDistList.append(0)
            else:
                if cnt < ptCount:
                    pDistList.append(PointDist(firstpt,(pointList[cnt - 1])))
```

```
            cnt += 1

        cnt = 0
        with arcpy.da.UpdateCursor(inGPSFC,["DistFromLastPoint"]) as UCursor:
            cnt = 0
            for uRow in UCursor:
                uRow[0] = pDistList[cnt]
                UCursor.updateRow(uRow)
                cnt += 1
        return inGPSFC
    except (RuntimeError, TypeError, NameError) as err:
        print("Oops! Error in the RefineTravel Function: " + err)
        return False
```

Appendix E: R Script of Random Forest Classifier

```
### Step 1: Load and initialize the arcgisbinding, random forest, and caret
### packages

library(arcgisbinding)
arc.check_product()
#install.packages('randomForest')
library(randomForest)
#install.packages('caret')
library(caret)


### Step 2: Load all FGDB Table (that drives the loop) containing the paths
### and naming structure into R.
### The <STEAMSeason>_FCPaths Table contains the list of all the GPS Feature
### Classes to process in R and the paths and names of output data from R

# Open the table
FCpaths <- arc.open(path =
'C:\\w\\PhDSTEAMProcessing\\Indoor_Outdoor\\DataPathsForR.gdb\\S2010_FCPaths')

# Select all the records
d_path <- arc.select(FCpaths, c('SID',
'Season','dPath','ErrorCSVPath','ForestSizeInfluencePlotPath','VariableImporta
ncePlotPath','ConfusionMatrixCSVPath','tPath'))

# Create an empty data frame with column names
edf <- data.frame( "SID" = character(0), "Season" = character(0),
"TrainingSample" = numeric(0), "OOB" = numeric(0), "TrainAccuracy" =
numeric(0),stringsAsFactors = FALSE)


# Loop over rows of the <STEAMSeason>_FCPaths dataframe, using the pathnames
# to Feature Classes to drive the loop.
# Each GPS FGDB Feature Class is processed in the loop

for (row in 1:nrow(d_path)) {
  aSID <- d_path[row, "SID"]
  theSeason <- d_path[row, "Season"]
  dpath  <- d_path[row, "dPath"]
  errorpath  <- d_path[row, "ErrorCSVPath"]
  FSIPpath  <- d_path[row, "ForestSizeInfluencePlotPath"]
  VIPpath  <- d_path[row, "VariableImportancePlotPath"]
  CMcsvPath  <- d_path[row, "ConfusionMatrixCSVPath"]
  tpath  <- d_path[row, "tPath"]

  print(paste("The SID is", aSID))
  print(paste("The path is", dpath))
  print(paste("The Season is", theSeason))
  print(paste("The Forest Size Influence Plot Path is ", FSIPpath))
  print(paste("The Variable Importance Plot Path is ", VIPpath))
```

```r
  print(paste("The path to output table is", tpath))

  ### Step 3: Loads the GPS Feature Class as an R dataframe
  d <- arc.open(path = dpath)
  d_all <- arc.select(d, c('CKEY', 'HEIGHT','SPEED', 'PDOP','HDOP','VDOP',
'INDOOR', 'DistFromLastPoint', 'DistFromBuild'), where_clause = "InCity = 1")
  d_CKey <- arc.select(d, c('CKEY'), where_clause = "InCity = 1")

  ### Step 4: Select the subset of known indoor/outdoor data to be used as the
  ### training and test dataset
  df <- arc.select(d, c('HEIGHT','SPEED','PDOP','HDOP','VDOP','INDOOR',
'DistFromLastPoint', 'DistFromBuild'), where_clause = "TrainingFlag = 1 AND
InCity = 1")

  # GET COUNT FROM TRAINING AND GENERATE SAMPLE SIZE
  indsample <- sample(2, nrow(df), replace = TRUE, prob = c(0.66,0.33))

  # 66% of the data flagged as "Training" will be used to TRAIN the random
  # forest model
  train_data <- df[indsample==1,]

  # 33% of the data flagged as "Training" will be used to TEST the random
  # forest model
  test_data <- df[indsample==2,]

  print(paste("Training Data Size:"))
  TrainSize <- nrow(train_data)
  print(TrainSize)

  print(paste("Test Data Size:"))
  print(nrow(test_data))

  outdoorTrainCnt <- nrow(subset(train_data, INDOOR < 1))
  print(paste("Number of Outdoor Pts in Training Data:", outdoorTrainCnt))
  outdoorTestCnt <- nrow(subset(test_data, INDOOR < 1))
  print(paste("Number of Outdoor Pts in Test Data:",outdoorTestCnt))

  # Only create an individual Random Forest Model for those participants with
  # at least 50 minute of a Training Sample and outdoors for 10 minutes
  # (Training Sample) and 5 minutes (Test Sample)
  # Else use the previous Random Forest Model from the last participant who
  # met these thresholds.
  if ((TrainSize > 3000) & (outdoorTrainCnt > 600) & (outdoorTestCnt > 300)) {
    print(paste("Running Random Forest Model Generator - Normal Mode"))

    ### Create the Random Forest Model
    rf_model <-
      randomForest(as.factor(INDOOR) ~ .,
                   train_data,
                   ntree = 500,
                   importance = TRUE)
    print(paste("Random Forest Model created using training data"))
    print(paste("Random Forest model - Confusion Matrix of Out-of-Bag and
          remaining Training observations:"))
    print(rf_model$confusion)
```

```
print(paste("Random Forest model: Num var at each split"))
print(rf_model$mtry)

# Plot the model to see the influence of Forest Size
# Open a pdf file
pdf(FSIPpath)
layout(matrix(c(1, 2), nrow = 1),
       width = c(4, 1))
par(mar = c(5, 4, 4, 0)) # No margin on the right side
plot(rf_model, log = "y")
par(mar = c(5, 0, 4, 2)) # No margin on the left side
plot(
  c(0, 1),
  type = "n",
  axes = F,
  xlab = "",
  ylab = ""
)
legend(
  "top",
  colnames(rf_model$err.rate),
  col = 1:4,
  cex = 0.8,
  fill = 1:4
)
dev.off()

# Gather the Model OOB Error
oobError <-
  mean(predict(rf_model) != as.factor(train_data$INDOOR))
print(paste("Random Forest Model: Training OOB Error"))
print(oobError)

# Run Training Data through the Random Forest Model
train_predict <- predict(rf_model, train_data)

# Step 5: Run Test Dataset through the Random Forest Model, create
# Confusion Matrix (.csv file), Model Accuracy and generate the variable
# importance plot to illustrate the variables were most important for
# training the model

print(paste("Step 5 train predict"))
test_data_p_model <- predict(rf_model, test_data, type = 'response')

print(paste("Confusion Matrix of Prediction and Test data sample:"))
cm <-
  confusionMatrix(test_data_p_model,
                  as.factor(test_data$INDOOR), positive =
                    "1")
trainAccuracy <- cm$overall['Accuracy']
tocsv <- cm$table
write.csv(tocsv, file = CMcsvPath)

# Create Variable Importance Plot
pdf(VIPpath)
```

```
      varImpPlot(rf_model)
      dev.off()

      ### Step 6-option 1: Run Entire Data Test Dataset through its own
      ### Random Forest Model
      print(paste("Prediction on full set of observations"))
      outin_predicted <- predict(rf_model, d_all, type = 'response')

      ### Step 8-option 1: Join the predicted Indoors/Outdoors class to the
      ### original GPS dataframe and output as a File Geodatabase Table in Arc

      print(paste("Cbind of prediction and full set of observations"))
      dc <- cbind(d_CKey, outin_predicted, deparse.level = 1)

      print(paste("Output Table to Arc"))
      arc.write(tpath, dc)

      ### STEP 7-option 1: Store the OOB Error Rate, training size, accuracy
      ### of the model
      edf[nrow(edf) + 1, ] = list(
        SID = aSID,
        Season = theSeason,
        TrainingSample = TrainSize,
        OOB = oobError,
        TrainAccuracy = trainAccuracy
      )
      lastSID = aSID

  } else {
      print(paste("SKIPPING Random Forest Model for DATASET ", sid, " using
Random Forest from : " ,lastSID))

      ### Step 6-option 2: Run Entire Data Test Dataset through the last
      ### valid Random Forest Model (the last one that contained a large enough
      ### training and test sample size

      print(paste("Prediction on full set of observations"))
      outin_predicted <- predict(rf_model, d_all, type = 'response')

      ### Step 8-option 2: Join the predicted Indoors/outdoors to the
      ### original GPS dataframe and output as a File Geodatabase Table in Arc
      print(paste("Cbind of prediction and full set of observations"))
      dc <- cbind(d_CKey, outin_predicted, deparse.level = 1)

      print(paste("Output Table to Arc"))
      arc.write(tpath, dc)

      ### STEP 7-option 2: Store the OOB Error Rate, training size, accuracy
      ### of the previous valid model

      print(paste("Appending to the GPS RF Stats Data Frame"))
      edf[nrow(edf) + 1, ] = list(
        SID = aSID,
        Season = theSeason,
        TrainingSample = TrainSize,
```

```
        OOB = oobError,
        TrainAccuracy = trainAccuracy
      )
    }
  }
#}
```

```
### STEP 8 – write out random Forest error and classification accuracy Stats
for the entire ### Run
print(paste("Completed Loop - now witing out GPS RF Stats"))
write.csv(edf,file=errorpath)
```

Appendix F: Post-Processing Script for Random Forest

```python
###=============================
### Program: Post-Process STEAM GPS Data from Random Forest model in R
### Author: Martin Healy
### Date: July 18, 2018
### Description: This program is designed to run from an ArcGIS Pro Script
###              Tool and prepares the data for use in SPSS:
###      1. Joins the Prediction Table from Random Forest to STEAM Feature
###       Classes
###      2. SID HID DayType FULLTIME Predicted exported to a csv suitable for
###       SPSS
###      3. Merge into one CSV
###
### Dependencies: arcpy, sys, os, gc, SteamModules
###=============================
import arcpy, csv, sys, os, gc

#FGDB = arcpy.GetParameterAsText(0)
FGDB = "C:\\w\\PhDSTEAMProcessing\\Indoor_Outdoor\\W2010.gdb"

#arcpy.env.workspace = arcpy.GetParameterAsText(1)
arcpy.env.workspace =
"C:\\w\\PhDSTEAMProcessing\\Indoor_Outdoor\\W2010.gdb"#"C:\\w\\PhDSTEAMProcess
ing\\Indoor_Outdoor\\S2010.gdb"

# Create and Open the output CSV file to contain the GPS records to Import to
# SPSS
output=open(r'C:\w\PhDSTEAMProcessing\Indoor_Outdoor\DataAnalysis\CSVFilesForS
PSS\W2010.csv','w',newline='')
linewriter=csv.writer(output,delimiter=',')
linewriter.writerow(["SID", "SEASON","HID", "DayType", "FULLTIME",
"outin_predicted"])
name_field = "InCity"

for tabl in arcpy.ListTables("*inout"):
    strLeng = len(tabl)
    FC = tabl[0:(strLeng - 5)]
    print("Processing: " + FC)
    arcpy.JoinField_management(FC,"CKEY",tabl,"CKEY",["outin_predicted"])
    expression = arcpy.AddFieldDelimiters(FC, name_field) + ' = 1'
    Scursor = arcpy.da.SearchCursor(FC,["SID", "Season","HID", "DayType",
                "FULLTIME", "outin_predicted"], where_clause=expression)
    for aRow in Scursor:
        sid = aRow[0]
        season = aRow[1]
        hid = aRow[2]
        daytype = aRow[3]
        ft = aRow[4]
        strFTime = ft.strftime("%m/%d/%Y %H%M%S %p")
        out_in = aRow[5]
        linewriter.writerow([sid, season, hid, daytype, strFTime, out_in])
```

Curriculum Vitae

Name: Martin A. Healy

Post-secondary   The University of Western Ontario
Education and   London, Ontario, Canada
Degrees: 1989 B.A.

                         College of Geographic Sciences
                         Lawrencetown, Nova Scotia, Canada
                         1994 Graduate Certificate, Geographic Information Systems

                         The University of Western Ontario
                         London, Ontario, Canada
                         2002 B.A. (Hons.)

                         The University of Western Ontario
                         London, Ontario, Canada
                         2007 M.A.

                         The University of Western Ontario
                         London, Ontario, Canada
                         2018 Ph.D.

Honours and   Gold Medal Award for highest overall grade in
Awards: undergraduate B.A. (Hons.) class (2003)

                         Canadian Association of Geographers for highest grade in
                         undergraduate thesis (2003)

Related Work   Faculty
Experience   School of Art and Design
                         Fanshawe College
                         2001-Present

Publications:

Gilliland, J. A., Rangel, C. Y., Healy, M. A., Tucker, P., Loebach, J. E., Hess, P. M., . . . Wilk, P. (2012). Linking childhood obesity to the built environment: a multi-level analysis of home and school neighbourhood factors associated with body mass index. Can J Public Health, 103(9 Suppl 3), eS15-21.

Healy, M. A., & Gilliland, J. A. (2012). Quantifying the magnitude of environmental exposure misclassification when using imprecise address proxies in public health research. Spat Spatiotemporal Epidemiol, 3(1), 55-67. doi:10.1016/j.sste.2012.02.006

Malczewski, J., Chapman, T., Flegel, C., Walters, D., Shrubsole, D., & Healy, M. A. (2003). GIS - multicriteria evaluation with ordered weighted averaging (OWA): case study of developing watershed management strategies. Environment and Planning A, 35(10), 1769-1784.