9-27-2018 12:00 PM

Space-Time Analysis of Breast Cancer in Middlesex County Between 2003 and 2013

Jenny T. Tjhin, *The University of Western Ontario*

Supervisor: Luginaah, Isaac, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Geography

Abstract

Breast cancer is the leading cancer among women in Canada and the incidence rate continues to increase despite improved screening programs and advanced cancer treatments. Some studies have linked breast cancer with risk factors such as genetics, the age of first menstruation, parity, and environmental exposures. This study investigated the spatial distribution of breast cancer cases between 2003 and 2013 in Middlesex County using geospatial techniques. Point data were analyzed with SaTScan and aggregated data were observed at two geographical units, census sub-divisions and dissemination areas using Moran's Index to detect clusters. Both analyses showed consistency in the cluster locations mostly in the western and eastern parts of the county. Age-adjusted breast cancer rates were then analyzed using multivariate and principal component analysis to explore potential links to socioeconomic factors obtained from the Canadian Census. Average income, employment rate, and occupations related to agriculture, forestry, fishing and hunting were significantly associated with increased breast cancer risk in the area. The findings provide pointers for local level policy related to breast cancer prevention and management in Middlesex County and beyond.

Keywords: breast cancer, space-time analysis, cluster analysis, Middlesex County

Acknowledgments

This thesis would not have been completed without the patience and guidance of my research supervisor, Dr. Isaac Luginaah. Thank you for challenging me with insightful questions and offering invaluable advice over the course of my graduate training.

I would like to acknowledge all the staff and faculty for teaching and guiding me throughout my coursework, TA-ship, and thesis. An exceptional mention goes to Lori Johnson whose reliability and bright personality never fail to keep me on top of the administrative side of this journey.

A special thank you goes to Dr. Benjamin Rubin who introduced me to the software *R* and thoroughly answered my many questions about statistics.

I am grateful for Martin Healy who taught me the fundamentals of spatial analysis and continues to support me in my journey.

I would like to extend my gratitude to Dr. Micha Pazner for his precious guidance.

Thank you to the former and current graduate students in the Environmental Hazards and Health Lab for your support.

Lastly, for their abundant support and inspiration, a heartfelt thank you is dedicated to my family and friends: Fany Tania, Rosalind Ragetlie, Mark Langton, Francine Visser, Alexandra Ouedraogo, Amy McKinnon, Jason Mulimba Were, Faraj Haddad, Ethan Shrubsole, Marilyn Smith, Nicole Moore, and Rimon Shalash.

Contents

List of Tables

List of Figures

1 Introduction

This thesis focuses on breast cancer cluster detection in Middlesex County, Ontario, Canada. The combination of patients' information from the cancer registry and population data from the census were examined using spatial analysis methods. The intent of the study is to identify the locations of clusters with a high risk of breast cancer in the county and explore potential socioeconomic factors that may increase the risk of breast cancer development for women in the study area.

This chapter begins by providing a brief explanation of breast cancer and its contextual background. The reader is then introduced to health geography in terms of the link between health and spatial analysis. The research questions and objectives are discussed and the chapter concludes with a complete overview of the thesis organization.

1.1 Breast cancer

The breast is a part of the reproductive system in women that is located on the chest and mostly made of fat, to serve as a mammary gland to produce milk for infants (National Cancer Institute, 2013). Cells in the breast undergo several changes in a woman's lifetime, starting with the initial development of the organ before puberty, regular changes during menstruation cycles, major changes during pregnancy, and the termination of hormones production during menopause (Johns Hopkins Medicine Health Library, 2013). These changes are regulated by two female hormones produced in the ovaries: estrogen and progesterone. Estrogen is mostly responsible for determining female sexual characteristics including breast development and menstrual processes, while progesterone controls milk production and the preparation of the womb for pregnancy.

According to the Canadian Cancer Society (CCS), breast cells that undergo unusual changes can grow into a tumour or a lump that begins to invade and destroy the surround-

ing tissues rendering these cells cancerous. Some cells may then break away from the original location and spread to other body parts (CCS, 2018). This results in breast cancer which can cause death if not treated or detected early enough. This disease mostly affects women who are 50 and older but also affects their younger counterparts. Over the years, breast cancer has drawn significant attention in population health research.

1.2 Population and health

In their 2014 Global Health Estimates report, the World Health Organization (WHO) reported breast cancer as the most common cancer in women, with varying incidence rates worldwide. The WHO indicated that more than half a million women died due to breast cancer in 2011 alone (WHO, 2014). In Canada, breast cancer is also the leading cancer for women, and it is expected that 1 in 8 will develop the disease in their lifetime (Statistics Canada, 2017).

Even though the breast cancer mortality rate in Canada is currently at its lowest level since 1950 due to interventions including early detection, regular screening, and better treatments, the incidence rate is still steadily increasing every year (CCS, 2018). In 2017, breast cancer accounted for 25% of all new cancer cases for women in the country (CCS, 2017), with more than half of them occurring in women between 50 and 69. The highest number of deaths caused by breast cancer occurred amongst women who were 80 years or older. Breast cancer affecting women under 50 generally tends to be more aggressive (Azim & Partridge, 2014). Hence, it is important to raise awareness of breast cancer for this age group and all women in public health programming.

The causes of breast cancer have been inconclusively reported due to high dependency on the geographical locations and population characteristics including education, income level, parity, use of screening services, breastfeeding, lifestyle and diet (Fejerman & Ziv, 2008). Furthermore, population dynamics may contribute to breast cancer prevention and treatment. For example, people who live in rural areas may have less access to breast screening programs

(Pong et al., 2009). Also, contextual characteristics including certain sociocultural beliefs may affect treatment seeking behaviours and such belief systems may be influenced by other social determinants of health (Mitchell et al., 2002). From a geographical point of view, certain locations may show an increased risk of breast cancer because of environmental risk factors including environmental pollution and other exposures (Dummer, 2008). For instance, breast cancer incidence has been reported to increase with traffic pollution in urban areas (Mordukhovich et al., 2016) and also with exposure to chemicals used in agricultural production (Reynolds et al., 2005; Brophy et al., 2006). Geospatial analysis can help to identify the locational disparities in breast cancer incidence to potentially explain risk factors and essentially, pave the way for improved health care policy and earlier intervention.

1.3 Spatial analysis in health

In health-related studies, the availability of geographically recorded health and population data has greatly assisted in the investigations of spatial patterns of disease (Elliott et al., 2001). Studies in population health that use spatial analysis tools are usually constructed with a combination of three processes (see Conceptual Framework section in Chapter 2). The first process involves disease mapping, where health problems are displayed using maps to draw conclusions visually. This exploratory and descriptive process involves different techniques that are used to represent clear and concise data and at the same time display interesting and informative maps. Colour management, human psychology in map reading, and map aesthetics are put together to produce maps that can highlight health problems.

Once data is visualized, exploratory analysis can be conducted to summarize the main characteristics. There are a large number of spatial analysis tools available for various contexts and purposes, such as measuring the distance between points, aggregating health data points into areas, calculating shortest distance from a residence to a hospital, measuring coverage of health services by distance, and many more. A concrete example that uses this process is

the annual map published by Public Health Ontario (PHO) displaying estimated risk areas for Lyme disease in the province, allowing the public to access this information and take extra precautions within the marked areas (PHO, 2018). These maps were generated from locations where black-legged ticks were spotted and, based on the estimation of ticks' behaviours in travelling and spreading Lyme disease, the analysis displayed areas within 20 km radius from the identified locations.

Taking spatial analysis one step further, studies have used various statistical methods to model the determinants of health. These types of studies investigate the correlation between health data locations and disease risk factors. When an *a priori* hypothesis exists, the researcher could examine spatial patterns based on suspected locations of a health hazard. For example, if there are complaints about water pollution near an industrial site, then the analysis will be driven by the location of that site as the centre of observation. In other cases where there is lack of etiological hypothesis, cluster detection can be performed using point pattern analyses or aggregated pattern trend analyses. The latter would face more challenges in interpretation, but robust statistical methods may help to reduce noise in the data and potentially highlight important etiological clues. These processes can overlap with each other and can be combined to explain health-related phenomena.

1.4 Geographical Information System (GIS)

Exploring environmental exposures in an area can help to better understand breast cancer etiology within a population. When looking at the geographic context of places and how places connect (Dummer, 2008), the technology of GIS may bring in potential environmental factors that contribute to the increased risk of breast cancer. As stated by Bailey and Gatrell '*space can make a difference*', whereby the use of spatial data may produce more meaningful outcomes than the analyses without spatial dimension (Bailey & Gatrell, 1995).

In an attempt to define GIS, it is important to take into account when the use for GIS was

first recognized and how it has grown in the last few decades. John Snow's cholera outbreak mapping in London, England in 1854 was the first known case that used a form of spatial analysis as a problem-solving tool. After gathering information from local residents, he began his study with a hypothesis that the cholera outbreak was related to the public water pumps. He drew a map of points to depict the locations of cholera cases as well as the water sources. The map showed that cholera cases formed clusters near a single pump that was contaminated. He suggested the local authority to disable this particular pump and this action ended the outbreak. This finding was a major cornerstone that connected geography and public health safety using spatial analysis.

Historically, maps have been created and used for the purpose of navigation and planning long before GIS was introduced, yet conducting a spatial analysis was difficult then because the data were often inaccurate, the distance measurement was cumbersome and area calculation was difficult. Between 1960 and 1980 the idea to transfer paper maps to a digital format for better computing became a reality, supported by the continuous advancement of computers' data storage. Digital maps and GIS were initially utilized at the government level for land planning and decision making in Canada and for census analysis in the United States. In the 1980s as well, the Ordnance Survey in the UK started to digitize detailed topographic maps for the whole country.

The term GIS was first mentioned by Roger Tomlinson (1966) when he was developing the Canada Geographical Information System (CGIS), the first computerized GIS in the world that was used to combine land use mapping and the emerging computer technology. His work was so brilliant that he is known today as 'the father of GIS'.

GIS has been defined as either a Geographical Information System or Science. As a system, GIS serves as a tool to capture, store, check, manipulate, analyze, and display data that are referenced to the Earth (Department of the Environment, 1987). In practice, GIS is not only used to explore objects on earth, but also the surfaces of other planets (Bell et al., 2007; Nass et al., 2011; Besse et al., 2017). From the science perspective, GIS is cultivated

by various concepts and ideas from many disciplines, including cartography, cognitive science, computer science, engineering, environmental sciences, law, and many more (Heywood et al., 2011). Goodchild et al. (1997) summarized the two mainstream definitions into an idea that Geographic Information Science is the science behind GIS technology.

According to the Environmental Systems Research Institute (Esri), the leading international company that develops comprehensive GIS applications, GIS is a framework where spatial data is stored, managed, and analyzed in a way that layers of information are organized to yield insights about the data to help in decision-making (Esri, 2017). Data can be managed using logical workflows, or altered with comprehensive mathematical tools to perform geographically appropriate analyses, or explored with various statistical methods to describe spatial patterns or make future predictions. GIS has been widely used to offer solutions for complex problems.

GIS continued to evolve after Tomlinson's breakthrough, with the combination of commercialization of various software, cheaper and faster computers, the launch of new satellites, remote sensing technology, and data availability. All of these factors contribute to GIS' rapid growth as a robust spatial analysis tool. This growth has made the application of GIS broader as it adjusts to different fields including government, defence, transportation, service planning, urban management, commerce and business, communications, environmental management, and health (Heywood et al., 2011).

1.5 Research questions

GIS tools, together with spatial analysis, have been employed to understand complex environmental and health phenomena. In the case of breast cancer, an understanding of its distribution over space and time would aid local level policies in cancer prevention and treatment programs. Therefore, this study examines the spatial distribution of breast cancer in Middlesex county and will answer the following research questions:

1. Are there any clusters of breast cancer in Middlesex County and its surrounding areas? If yes, where are they located?

2. Are there any socioeconomic factors that contribute to the elevated risk of breast cancer within the identified clusters?

1.6 Organization of thesis

The first chapter introduces the thesis topics including breast cancer burden to the population in Canada, the history of the definition of GIS, the role of spatial analysis in epidemiology, and the statement of research questions that drive the research. Chapter 2 states the research problem in detail and presents relevant published literature that explains how other studies tried to identify breast cancer risk using spatial analysis and GIS tools. The gap in the literature is identified here and the important issues that need investigation are listed.

Chapter 3 is a manuscript that reports details of the study which contains data sources, methods used, and the outcomes of the research. Finally, Chapter 4 concludes the research and offers directions for future research that may be applied to better understand the etiology of breast cancer.

There are two appendices that list socioeconomic variables used in the analysis and parameters for the software to perform cluster detection. Since this is a manuscript thesis, there are some areas that may be repetitive.

References

Azim, H. A., & Partridge, A. H. (2014, aug). Biology of breast cancer in young women. *Breast cancer research : BCR*, *16*(4), 427. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25436920http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4303229 doi: 10.1186/s13058-014-0427-5

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Longman Scientific & Technical. Retrieved from https://books.google.ca/books/about/Interactive{_}spatial{_}data{_}analysis.html?id=WwbvAAAAMAAJ

Bell, D. G., Kuehnel, F., Maxwell, C., Kim, R., Kasraie, K., Gaskins, T., ... Coughlan, J. (2007). NASA World Wind: Opensource GIS for Mission Operations. In *2007 ieee aerospace conference* (pp. 1–9). IEEE. Retrieved from http://ieeexplore.ieee.org/document/4161692/ doi: 10.1109/AERO.2007.352954

Besse, S., Vallat, C., Barbarisi, I., Arviset, C., De Marchi, G., Barthelemy, M., ... Saiz, J. (2017). *The New Planetary Science Archieve (PSA): Exploration and discovery of scientific datasets from ESA's planetary missions* (Tech. Rep.). Retrieved from http://psa.esa.int

Brophy, J. T., Keith, M. M., Gorey, K. M., Luginaah, I., Laukkanen, E., Hellyer, D., ... Gilbertson, M. (2006). Occupation and breast cancer: A Canadian case-control study. *Annals of the New York Academy of Sciences*, *1076*, 765–777. doi: 10.1196/annals.1371.019

Canadian Cancer Society. (2017). *Breast cancer statistics - Canadian Cancer Society.* Retrieved 2017-01-30, from http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on

Canadian Cancer Society. (2018). *Cancer clusters.* Retrieved 2018-05-05, from http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/cancer-clusters/?region=on

Department of the Environment. (1987). *Handling Geographic Information* (Tech. Rep.). London, Her Majesty's Stationery Office. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/02693798708927805

Dummer, T. J. B. (2008, apr). Health geography: supporting public health policy and planning. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, *178*(9), 1177–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18427094http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2292766 doi: 10.1503/cmaj.071783

Elliott, P., Wakefield, J., Best, N., & Briggs, D. (2001). *Spatial Epidemiology Methods and Applications*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198515326.001.0001/acprof-9780198515326 doi: 10.1093/acprof:oso/9780198515326.001.0001

Esri. (2017). *ArcGIS Desktop.* Redlands, CA: Environmental Systems Research Institute.

Fejerman, L., & Ziv, E. (2008, mar). Population differences in breast cancer severity. *Pharmacogenomics*, *9*(3), 323–333. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18303968https://www.futuremedicine.com/doi/10.2217/14622416.9.3.323 doi: 10.2217/14622416.9.3.323

Goodchild, M., & Quattrochi, D. (1997). *Scale, multiscaling, remote sensing, and GIS*. Retrieved from http://www.citeulike.org/group/7954/article/4257773

Heywood, D., Cornelius, S., & Carver, S. (2011). *An introduction to geographical information systems*.

Johns Hopkins Medicine Health Library. (2013). *Normal Breast Development and Changes —*. Retrieved 2018-06-27, from http://www.hopkinsmedicine.org/healthlibrary/conditions/breast{_}health/normal{_}breast{_}development{_}and{_}changes{_}85,P00151/

Mitchell, J., Lannin, D. R., Mathews, H. F., & Swanson, M. S. (2002, dec). Religious Beliefs and Breast Cancer Screening. *Journal of Women's Health*, *11*(10), 907–915. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12626089http://www.liebertonline.com/doi/abs/10.1089/154099902762203740 doi: 10.1089/154099902762203740

Mordukhovich, I., Beyea, J., Herring, A. H., Hatch, M., Stellman, S. D., Teitelbaum, S. L., ... Gammon, M. D. (2016, jan). Vehicular Traffic-Related Polycyclic Aromatic Hydrocarbon Exposure and Breast Cancer Incidence: The Long Island Breast Cancer Study Project (LIBCSP). *Environmental health perspectives*, *124*(1), 30–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26008800http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4710589 doi: 10.1289/ehp.1307736

Nass, A., van Gasselt, S., Jaumann, R., & Asche, H. (2011, sep). Implementation of cartographic symbols for planetary mapping in geographic information systems. *Planetary and Space Science*, *59*(11-12), 1255–1264. Retrieved from https://www.sciencedirect.com/science/article/pii/S0032063310002606 doi: 10.1016/J.PSS.2010.08.022

National Cancer Institute. (2013). *NCI Dictionary of Cancer Terms* (Vol. 1) (No. September 2013). Retrieved 2018-06-27, from https://www.cancer.gov/publications/dictionaries/cancer-terms/search?contains=false{\&}q=breast doi: 10.1017/CBO9781107415324.004

Pong, R. W., DesMeules, M., & Lagacé, C. (2009, feb). Rural-urban disparities in health: How does Canada fare and how does Canada compare with Australia? *Australian Journal of Rural Health*, *17*(1), 58–64. Retrieved from http://doi.wiley.com/10.1111/j.1440-1584.2008.01039.x doi: 10.1111/j.1440-1584.2008.01039.x

Public Health Ontario. (2018). *Ontario Lyme disease map 2018: estimated risk areas* (Tech. Rep.). Retrieved from https://www.publichealthontario.ca/en/eRepository/Lyme{_}disease{_}risk{_}areas{_}map.pdf

Reynolds, P., Hurley, S. E., Gunier, R. B., Yerabati, S., Quach, T., & Hertz, A. (2005, aug). Residential proximity to agricultural pesticide use and incidence of breast cancer in California, 1988-1997. *Environmental Health Perspectives*, *113*(8), 993–1000. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16079069http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1280339 doi: 10.1289/ehp.7765

Statistics Canada. (2017). *Census dictionary.* Retrieved 2017-11-05, from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

Tomlinson, R. F. (1966). A geographic information system for regional planning. *Stewart G A (ed) Land Evaluation*.

World Health Organization. (2014). *Global Health Estimates 2014: DALYs* (Tech. Rep.). Retrieved from http://www.who.int/healthinfo/global{_}burden{_}disease/en/

2 Research context

The scope of the study is discussed in this chapter. It starts with a profile of Middlesex County that will inform the reader about the characteristics of the area. Afterwards, this section describes the local health unit who has agreed to collaborate with us to mobilize the knowledge from this study. Ideas regarding the thesis approach to achieve our research objectives are then discussed in the conceptual framework section. Finally, the chapter concludes the research context with a literature review of previous breast cancer studies that utilize spatial analysis.

2.1 Community profile

Middlesex County is a mix of rural and urban areas in Ontario, Canada that covers 3,318 km^2 with a population of 439,151 according to the 2011 census. The county population grew by 4% between 2011 and 2016 (Statistics Canada, 2011). It is land-locked by seven counties including Huron to the north, followed clockwise by Perth, Oxford, Elgin-St. Thomas, Chatham-Kent, and Lambton (Figure 3.1a). The population distribution in the area varies based on urban and rural concentrations. Sitting at the south-centre area of the county along the major highways corridor, is the City of London, which serves as the administrative capital. Located in the heart of Southwestern Ontario with vibrant downtown cores, commercial plazas, and industrial lands, the Middlesex County is suitable for commercial and industrial development. Some of the rural areas are the most fertile in the province and have enabled the agricultural sector to thrive (Middlesex Economic Development, 2018).

Each county or Census Divisions (CD) in the province of Ontario is divided into Census Sub-Divisions (CSD) that represent municipalities or areas including Indian reserves, Indian settlements, and unknown territories. Middlesex County is composed of eight local municipalities including Adelaide Metcalfe, Lucan Biddulph, Middlesex Centre, North Middlesex, Southwest Middlesex, Strathroy-Caradoc, Thames Centre, and Village of Newbury, and three

First Nations communities: Chippewas of the Thames First Nation, Munsee-Delaware Nation and Oneida. CSDs are then divided into smaller administrative areas, which are Census Tracts (CT), and each CT is subdivided into several Dissemination Areas (DA). The majority of CSDs have less than 1,000 people while each DA is populated by between 400 and 700 dwellers (Statistics Canada, 2017). We use spatial analysis techniques to examine the incidence of breast cancer in the county over space and time.

2.2 Conceptual framework

This study is ecological with respect to breast cancer cases and socioeconomic factor measurements. Observing health in an aggregated form of population involves inherent dynamics of the population that require careful considerations including place of birth, daily movements for work commutes, considerable time spent for various interests, and population migration that exposes each individual to different environmental factors. Moreover, individuals have varying age, gender, genetic factors, and choices of lifestyles that may obscure important details when those factors are combined to search for trends in population health data.

Ideally, a complete and precise record of individuals would need to be collected and analyzed to produce high-quality results that depict health problems in the population, so that interventions may reach a precise target to completely mitigate health risks. But in reality, the richness of health data is more constrained within boundaries of ethics, personal preferences, confidentiality, and socially constructed paradigms. Health studies that are carried out using reasonably good quality data within small areas can better target health interventions because it focuses on specific issues that apply to that region.

Figure 2.1 shows an interconnectivity of different elements of spatial analysis that build a conceptual framework for this study. It is a modification of the framework of spatial epidemiological data analysis suggested by Pfeiffer et al. (2008). Various techniques are available to implement spatial analysis but the main analyses in this study are categorized into the descrip-

tion of spatial patterns (visualization), cluster detection (exploration), and the explanation of disease risk (modelling).

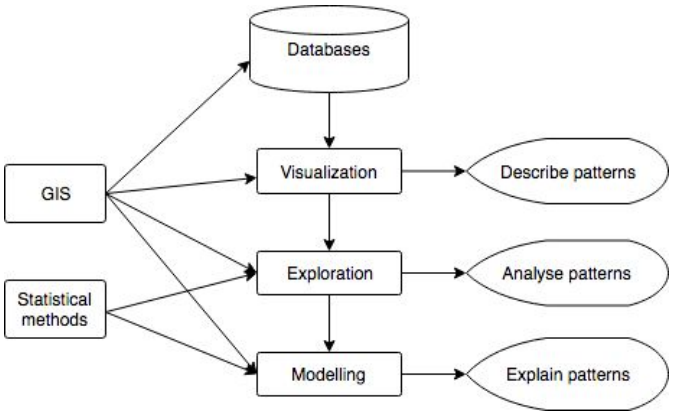

Figure 2.1: Conceptual framework of the study adopted from Pfeiffer et al. (2008)

The objectives of spatial analysis in health-related studies can be met with the availability of geographically referenced health data and other supporting datasets. The first two processes use spatial data to produce spatial products. The visualization can help to describe spatial patterns and communicate results, while the exploration involves statistical methods to detect clusters of disease and test whether the identified patterns happen due to chance. Lastly, the modelling, which may depart from exploring the concept of cause-effect relationships but not restricted to causal inference, may combine spatial and non-spatial data to explain or predict disease risks. GIS is the tool used to encompass the data storage and all of the processes for this study with the use of statistical methods for exploration and modelling.

2.3 Breast cancer studies

Breast cancer is the most frequent cancer diagnosis among women globally and North American rates are amongst the highest in the industrialized world (Ferlay et al., 2010). In the past few decades, the CCS reported that the total number of deaths caused by breast cancer has gradually increased at a steady rate in Canada (CCS, 2012). The report also mentioned that breast cancer was the leading cancer among Canadian women making up about a quarter of

all cancer cases. In 2017, it was predicted that one out of eight women would develop breast cancer in their lifetime.

Known risk factors of breast cancer include socioeconomic status, age, hormonal variation, and family history of breast cancer (Lynch et al., 1989; Trichopoulos et al., 1983; Ye et al., 2002; Lanfranchi, 2015). These known risk factors account for approximately 30% of the variation in breast cancer incidence (Fenga, 2016; Kamińska et al., 2015). It has been suggested that the remaining 70% of unexplained variance may be attributed to environmental exposure (e.g., contamination) (Fenga, 2016). Given the inherent challenges in establishing cause and effect relationships between environmental exposure and health outcomes however, research on the association between environmental exposures and breast cancer is conflicting. For example, Reynolds et al. found no increased risk of breast cancer with proximity to agricultural pesticide use, while Mills and Yang observed a positive association between organochlorine pesticides and breast cancer among Hispanics in California using county-level data (Reynolds et al., 2005; Mills & Yang, 2006).

In the Canadian context, studies have found that residential proximity to steel and pulp mills, thermal power plants, and petroleum refineries significantly increased the probability of breast cancer (Pan et al., 2011). Recent studies have also reported an association between the incidence of postmenopausal breast cancer and exposure to ambient concentrations of nitrogen dioxide and ultrafine particulate matter (Goldberg et al., 2017). Furthermore, previous work on the role of occupational exposure on breast cancer found elevated rates of breast cancer among women in jobs with potentially high exposures to carcinogens and endocrine disruptors (Soto & Sonnenschein, 2010; Macon & Fenton, 2013). In particular, the risk of breast cancer among women who had farmed was found to be significantly higher than women with no farming experience (Brophy et al., 2006). In Southern Ontario, numerous geographical areas associated with high rates of industrialization and trans-boundary air pollution present as sentinel areas of environmental carcinogen exposure(Green Brody et al., 2007). Additionally, this region has the most fertile soils in Canada and has therefore provided opportunities for a range of agricultural

applications that have, over the years, been associated with the use of a range of pesticides, herbicides and fungicides.

Several agencies and authors have called for the need to improve primary breast cancer prevention through a reduction of exposure to chemicals that may be potential carcinogens. Although exposure to environmental contaminants has received increasing attention, what has been ignored so far is the attempt to examine the role of environmental exposure to carcinogens on breast cancer incidence in Ontario.

Environmental exposure may cause disruption of hormones. In a meta-analysis article, Schneider et al. (2014) narrated comprehensive findings of hormone replacement therapy (HRT). HRT was initially introduced to help women reduce unpleasant menopausal symptoms by replacing estrogen that was no longer produced by the body. Many studies related the therapy to cardiovascular diseases and excessive occurrence of breast cancer amongst women who participated in the therapy. The health outcomes were so adverse that when the Women's Health Initiative made the news public, many women stopped the therapy and the treatment lost its appeal (Rossouw et al., 2002). Concomitantly, breast cancer rate went down significantly which suggested that the change of hormone balance was possibly related to an increased risk of breast cancer (Schneider et al., 2014). Furthermore, they reported that the incidence rate significantly increased in the 1990s due to the better access to mammography screening which resulted in a higher count of cases during those years. The rate fluctuated thereafter, except for a dramatic decrease of occurrence in 2002 when many women stopped using HRT.

Another piece of evidence to support the link between breast cancer and hormone disruption was published in the New England Journal of Medicine by Yager and Davidson (2006) regarding the use of oral contraceptives. They reported that the concentration of estrogen in breast tissues is much higher than in the rest of the body, and as such, the side effects of estrogen may lead to changes in the breasts. Other studies also reported an increased risk of breast cancer among oral contraceptive users (Kahlenborn et al., 2006; Hunter et al., 2010).

The timing, duration, and magnitude of occupational or environmental exposures are

valuable facts to understand breast cancer causality but are difficult to record and characterize (Schettler, 2013). One way to find associations between environmental exposures and an increased risk of disease is by observing the spatial pattern of incidents over a period of time. Marshall (1991) suggests that a disease can happen in clusters of space, or time, or both and that the concept of disease clustering is related to a statistically high incidence that is different from a gradual trend. In response, studies have been conducted to detect cancer clusters over a period of time in different geographical places. For instance, Kulldorff (1997) used the spatial scan statistics method to understand the spatial trends of breast cancer in New York. The same method was used to detect brain cancer clusters in Los Alamos, New Mexico (Kulldorff et al., 1998). Sherman et al. (2014) also looked at the early detection of colorectal cancer by finding locations of clusters which became the focus of public health interventions.

Research in breast cancer faces many challenges due to its association with interconnecting multiple risk factors, which have been reported inconclusive as a whole, yet may be specific to geographical locations. This study addresses breast cancer burden in the population and aims to better understand the etiology from spatial and statistical points of view. Geostatistical methods that have been widely used to assist in disease cluster detection include the Morans Index to determine spatial autocorrelation (Anselin, 1996) and Hotspot Analysis (Anselin, 1995). Spatial autocorrelation has been used to analyze cancer patterns in Western Europe (Rosenberg et al., 1990) and breast cancer patterns in Kentucky (Lin & Zhang, 2007). Hotspot analysis is a useful tool to find patterns of high occurrence or the prevalence of diseases.

Within the context of breast cancer in Ontario, a space-time analysis was conducted by Luginaah et al. (2012) between 1986 and 2002 using Kulldorff's method to find areas with high risk of breast cancer. This research extends Luginaah's study to understand the pattern of breast cancer in a different time period with a focus on a specific geographical location, Middlesex County. The need for this study is driven by the collaboration with the Middlesex-London Health Unit (MLHU) as they are interested in identifying areas in the county where they can focus on public health programs. Overall, this study is influenced by social determinants of

healthcare framework that expands our understanding of the multi-factors that can influence health directly or indirectly. Identifying breast cancer incident clusters over time can suggest associations to environmental contaminants or occupational risks related to the particular geographic areas.

References

Anselin, L. (1995, sep). Local Indicators of Spatial AssociationLISA. *Geographical Analysis*, *27*(2), 93–115. Retrieved from http://doi.wiley.com/10.1111/j.1538-4632.1995.tb00338.x doi: 10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (1996). The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In *Spatial analytical perspectives in gis* (pp. 111–125). Taylor & Francis. Retrieved from http://dces.wisc.edu/wp-content/uploads/sites/30/2013/08/W4{_}Anselin1996.pdf

Brophy, J. T., Keith, M. M., Gorey, K. M., Luginaah, I., Laukkanen, E., Hellyer, D., ... Gilbertson, M. (2006). Occupation and breast cancer: A Canadian case-control study. *Annals of the New York Academy of Sciences*, *1076*, 765–777. doi: 10.1196/annals.1371.019

Canadian Cancer Society. (2012). *Breast Cancer in Canada Detection problems.* Retrieved 2018-06-27, from http://www.cbcf.org/ontario/AboutBreastCancerMain/FactsStats/Pages/Breast-Cancer-Canada.aspx

Fenga, C. (2016). Occupational exposure and risk of breast cancer. *Biomedical Reports*, *4*, 282–292. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4774377/pdf/br-04-03-0282.pdf doi: 10.3892/br.2016.575

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010, dec). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, *127*(12), 2893–2917. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21351269http://doi.wiley.com/10.1002/ijc.25516 doi: 10.1002/ijc.25516

Goldberg, M. S., Labrèche, F., Weichenthal, S., Lavigne, E., Valois, M. F., Hatzopoulou, M., ... Parent, M. É. (2017, oct). The association between the incidence of postmenopausal breast cancer and concentrations at street-level of nitrogen dioxide and ultrafine particles. *Environmental Research*, *158*, 7–15. Retrieved from http://www.sciencedirect.com/science/article/pii/S0013935117304966 doi: 10.1016/j.envres.2017.05.038

Green Brody, J., Moysich, K. B., Humblet, O., Attfield, K. R., Beehler, G. P., & Rudel, R. A. (2007). Environmental Factors in Breast Cancer Environmental Pollutants and Breast Cancer Epidemiologic Studies. Retrieved from www.interscience.wiley.com doi: 10.1002/cncr.22655

Hunter, D. J., Colditz, G. A., Hankinson, S. E., Malspeis, S., Spiegelman, D., Chen, W., ... Willett, W. C. (2010, oct). Oral contraceptive use and breast cancer: a prospective study of young women. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *19*(10), 2496–502. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20802021http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3055790 doi: 10.1158/1055-9965.EPI-10-0747

Kahlenborn, C., Modugno, F., Potter, D. M., & Severs, W. B. (2006, oct). Oral Contraceptive Use as a Risk Factor for Premenopausal Breast Cancer: A Meta-analysis. *Mayo Clinic Proceedings*, *81*(10), 1290–1302. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17036554http://linkinghub.elsevier.com/retrieve/pii/S002561961161152X doi: 10.4065/81.10.1290

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., & Starosławska, E. (2015, sep). Breast cancer risk factors. *Przeglad menopauzalny = Menopause review*, *14*(3), 196–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26528110http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4612558 doi: 10.5114/pm.2015.54346

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*(6), 1481–1496. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03610929708831995 doi: 10.1080/03610929708831995

Kulldorff, M., Athas, W. F., Feuer, E. J., Miller, B. A., & Key, C. R. (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health*, *88*(9), 1377–1380. doi: 10.2105/AJPH.88.9.1377

Lanfranchi, A. (2015). Induced abortion and breast cancer. *Law & Medicine*, *30*(2), 143.

Lin, G., & Zhang, T. (2007, jul). Loglinear Residual Tests of Moran's I Autocorrelation and their Applications to Kentucky Breast Cancer Data. *Geographical Analysis*, *39*(3), 293–310. Retrieved from http://doi.wiley.com/10.1111/j.1538-4632.2007.00705.x doi: 10.1111/j.1538-4632.2007.00705.x

Luginaah, I. N., Gorey, K. M., Oiamo, T. H., Tang, K. X., Holowaty, E. J., Hamm, C., & Wright, F. C. (2012). A geographical analysis of breast cancer clustering in southern Ontario: Generating hypotheses on environmental influences. *International Journal of Environmental Health Research*, *22*(3), 232–248. doi: 10.1080/09603123.2011.634386

Lynch, H. T., Marcus, J. N., Watson, P., & Lynch, J. F. (1989). Familial and Genetic Factors New Evidence. In B. A. Stoll (Ed.), *Women at high risk to breast cancer* (pp. 27–39). Springer Netherlands. doi: 10.1007/978-94-009-1327-1_3

Macon, M. B., & Fenton, S. E. (2013, mar). Endocrine disruptors and the breast: early life effects and later life disease. *Journal of mammary gland biology and neoplasia*, *18*(1), 43–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23417729http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3682794 doi: 10.1007/s10911-013-9275-7

Marshall, R. J. (1991). A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease Author (s): Roger J . Marshall Source : Journal of the Royal Statistical Society . Series A (Statistics in Society), Vol . 154 , Published by : Wiley for the Royal Statist. *Journal of the Royal Statistical Society*, *154*(3), 421–441.

Middlesex Economic Development. (2018). *Our Community*. Retrieved 2018-06-30, from https://www.investinmiddlesex.ca/our-community

Mills, P. K., & Yang, R. (2006). *Regression Analysis of Pesticide Use and Breast Cancer Incidence in California Latinas* (Vol. 68) (No. 6). Retrieved 2017-11-02, from http://0-search.ebscohost.com.library.unl.edu/login.aspx?direct=true{\&}db=a2h{\&}AN=19570059{\&}site=ehost-live{\&}scope=site

Pan, S. Y., Morrison, H., Gibbons, L., Zhou, J., Wen, S. W., DesMeules, M., & Mao, Y. (2011, may). Breast Cancer Risk Associated With Residential Proximity to Industrial Plants in Canada. *Journal of Occupational and Environmental Medicine*, *53*(5), 522–529. Retrieved from http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage{\&}an=00043764-201105000-00010 doi: 10.1097/JOM.0b013e318216d0b3

Pfeiffer, D. (2008). *Spatial analysis in epidemiology*. Oxford University Press.

Reynolds, P., Hurley, S. E., Gunier, R. B., Yerabati, S., Quach, T., & Hertz, A. (2005, aug). Residential proximity to agricultural pesticide use and incidence of breast cancer in California, 1988-1997. *Environmental Health Perspectives*, *113*(8), 993–1000. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16079069http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1280339 doi: 10.1289/ehp.7765

Rosenberg, L., Palmer, J. R., Miller, D. R., Clarke, E. A., & Shapiro, S. (1990). A case-control study of alcoholic beverage consumption and breast cancer. *Am J Epidemiol*, *131*(1), 6–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2293754

Rossouw, J., Anderson, G., Prentice, R., LaCroix, A., Kooperbert, C., & Stefanick, M. (2002, jul). Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA: The Journal of the American Medical Association*, *288*(3), 321–333. Retrieved from http://jama.ama-assn.org/cgi/doi/10.1001/jama.288.3.321 doi: 10.1001/jama.288.3.321

Schettler, T. (2013). The Ecology of Breast Cancer. *Science & Environmental Health Network*(October), 1–200.

Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., & Windisch, A. K. (2014). *The breast cancer epidemic: 10 facts* (Vol. 81) (No. 3). Retrieved from http://www.tandfonline.com/doi/full/10.1179/2050854914Y.0000000027 doi: 10.1179/2050854914Y.0000000027

Sherman, R. L., Henry, K. a., Tannenbaum, S. L., Feaster, D. J., Kobetz, E., & Lee, D. J. (2014). Applying spatial analysis tools in public health: an example using SaTScan to detect geographic targets for colorectal cancer screening interventions. *Preventing chronic disease*, *11*(2), E41. Retrieved from http://www.pubmedcentral.nih.gov/articlerender

.fcgi?artid=3965324{\&}tool=pmcentrez{\&}rendertype=abstract doi: 10.5888/pcd11.130264

Soto, A. M., & Sonnenschein, C. (2010, jul). Environmental causes of cancer: endocrine disruptors as carcinogens. *Nature reviews. Endocrinology*, *6*(7), 363–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20498677http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3933258 doi: 10.1038/nrendo.2010.87

Statistics Canada. (2011). *2011 Census Program Data Products.*

Statistics Canada. (2017). *Census dictionary.* Retrieved 2017-11-05, from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

Trichopoulos, D., Hsieh, C.-C., Macmahon, B., Lln, T.-M., Lowe, C. R., Mirra, A. P., ... Yuasa, S. (1983). Age at any birth and breast cancer risk. *International Journal of Cancer*, *31*(6), 701–704. Retrieved from http://doi.wiley.com/10.1002/ijc.2910310604 doi: 10.1002/ijc.2910310604

Yager, J. D., & Davidson, N. E. (2006, jan). Estrogen Carcinogenesis in Breast Cancer. *New England Journal of Medicine*, *354*(3), 270–282. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16421368http://www.nejm.org/doi/abs/10.1056/NEJMra050776 doi: 10.1056/NEJMra050776

Ye, Z., Gao, D. L., Qin, Q., Ray, R. M., & Thomas, D. B. (2002). Breast cancer in relation to induced abortions in a cohort of Chinese women. *Br J Cancer*, *87*(9), 977–981. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12434288{\%}5Cnhttp://www.nature.com/bjc/journal/v87/n9/pdf/6600603a.pdf doi: 10.1038/sj.bjc.6600603

3 Space-Time Analysis of Breast Cancer in Middlesex County between 2003 and 2013

This chapter is a manuscript that consists of five sections that explain comprehensive details about the study. The sections are in the order of introduction, materials and methods, results, discussion, and conclusion. The introduction highlights the burden of breast cancer in our population, followed by an explanation about factors that increase the risk. A literature review is provided to show the background story of studies of breast cancer and spatial analysis. This story leads to an identification of a research gap this study attempts to fill. This section ends with the statement of the study objectives.

The second section describes the data and methods used to conduct analysis in the study. The first part of the analyses is examining the spatial distribution of breast cancer incidents to detect clusters. The second part aims to explore potential socioeconomic factors that may contribute to an increased risk of breast cancer in the identified clusters.

The results are reported in the third section and are discussed in Section 4. The study concludes by stating a summary of findings and recommendations for the local health unit.

3.1 Introduction

As the most prevalent cancer and the most common cause of cancer-related mortality among women, breast cancer burden is substantial in Canada (Statistics Canada, 2017). Breast cancer has been associated with many risk factors including genetics, the age of first menstruation, parity, age, hormonal variations, and family history (Lynch et al., 1989; Trichopoulos et al., 1983; Ye et al., 2002; Lanfranchi, 2015). These factors account for only about one-third of breast cancer variances and causes for the remaining cases are still widely unknown, but have been linked to environmental factors (e.g.: contamination) (Fenga, 2016; Kamińska et al., 2015). Studies that explore the link between breast cancer and environmental factors face many chal-

lenges due to their association with complex multiple risk factors that are inconclusive as a whole, yet may be specific to geographical locations. Timing, duration, and magnitude of the exposures are valuable facts needed to understand breast cancer causality but are difficult to record and characterize (Schettler, 2013).

Breast cancer has been linked to the use of organochlorines, which are synthetic chemicals released into the environment through the use of pesticides or industrial substances. Studies have reported conflicting results in this subject, for example, there was no association found linking these substances to breast cancer in North America and Europe (Calle et al., 2002; Reynolds et al., 2004), but a positive link between the two was found in later years (He et al., 2017; Mills & Yang, 2006). Furthermore, the risk of breast cancer has been associated with environmental factors including residential proximity to steel and pulp mills, thermal power plants, and petroleum refineries (Pan et al., 2011) and exposure to ambient concentrations of nitrogen dioxide and ultrafine particulate matter (Crouse et al., 2010). Breast cancer was also found to increase with occupational exposures to carcinogens and endocrine disruptors in Southern Ontario (Brophy et al., 2006). This region has the most fertile soils in Canada and has therefore provided opportunities for a range of agricultural applications of a range of pesticides, herbicides and fungicides over many years.

Exploring patterns of disease clusters according to their geographical areas may provide etiological clues; therefore, spatial analysis of breast cancer patterns can be used to identify potential environmental factors that may increase the risk and potentially better understand unexplained cases (Gatrell et al., 1996; Luginaah et al., 2012). Known disease clusters can suggest links to geographical, environmental, or occupational risks, providing a foundation for in-depth epidemiological investigations to find associations with environmental contaminants and associated carcinogens at or near where the clusters are located.

The purpose of this study is firstly to identify areas with higher than average risk of breast cancer that are considered clusters and secondly to explore possible socioeconomic factors that can increase that risk. This manuscript starts with a discussion of the data and methods that

are used in the study, followed by a presentation of results. A discussion of the findings in the context of breast cancer in the region is then presented, followed by the conclusion and study limitations.

3.2 Materials and Methods

3.2.1 Study Area

In 2011, Ontario was the most highly populated province in Canada with almost 40% of the total country's population (Statistics Canada, 2011). Cancer Care Ontario (CCO) reported that Middlesex County, which sits in the south-western part of the province, had a higher breast cancer incidence rate compared to the provincial rate between 2011 and 2013 (CCO, 2018). The county is a mix of rural and urban areas in Ontario, Canada that covers 3,318 km^2 with a population of 439,151 according to the 2011 census. The county population grew by 4% between 2011 and 2016 (Statistics Canada, 2011). It is land-locked by seven counties including Huron to the north, followed clockwise by Perth, Oxford, Elgin-St. Thomas, Chatham-Kent, and Lambton (Figure 3.1a). The population distribution in the area varies based on urban and rural concentrations. Sitting at the south-centre area of the county along the major highways corridor, is the City of London, which serves as the administrative capital. Located in the heart of Southwestern Ontario with vibrant downtown cores, commercial plazas, and industrial lands, the county is suitable for commercial and industrial development. Some of the rural areas are the most fertile in the province and have enabled the agricultural sector to thrive (Middlesex Economic Development, 2018).

Each county or Census Divisions (CD) in the province of Ontario is divided into Census Sub-Divisions (CSD) that represent municipalities or areas including Indian reserves, Indian settlements, and unknown territories. Middlesex County is composed of eight local municipalities including Adelaide Metcalfe, Lucan Biddulph, Middlesex Centre, North Middlesex, Southwest Middlesex, Strathroy-Caradoc, Thames Centre, and Village of Newbury, and three

First Nations communities: Chippewas of the Thames First Nation, Munsee-Delaware Nation and Oneida. CSDs are then divided into smaller administrative areas, which are Census Tracts (CT), and then each CT is divided into several Dissemination Areas (DA). The majority of CSDs have less than 1,000 people while each DA is populated by between 400 and 700 dwellers (Statistics Canada, 2017).

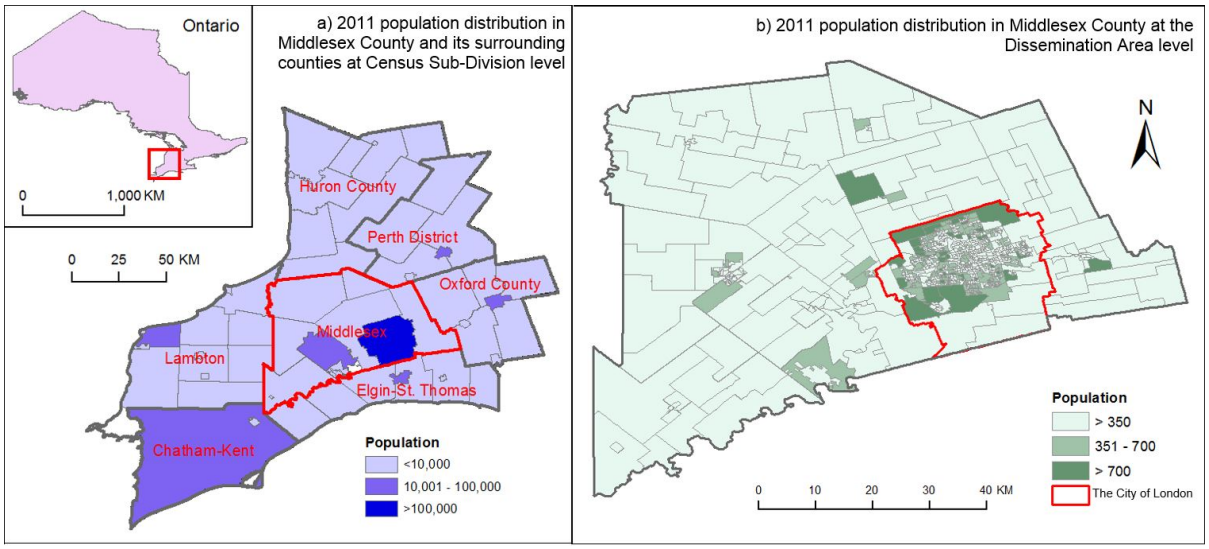

Figure 3.1: Maps of population distribution from 2011 census data

This study conducted spatial and temporal analysis at two levels of geographical units. Firstly, breast cancer prevalence was measured at the DA level (n=713) to identify clusters within the county. The DA was chosen because our cancer data and population data from the census database were available at this level and the granularity may increase the specificity of our results. Secondly, the study area was extended to include the county and its land-locking counties including Huron to the north followed clockwise by Perth, Oxford, Elgin-St. Thomas, Chatham-Kent, and Lambton, whose health units shared the same boundaries and the prevalences were measured at the CSD level (n=59).

Population distribution in the study area varied based on urban and rural concentrations. The top left map in Figure 3.1a shows the province of Ontario as a reference to our study area. The majority of CSDs have less than 1,000 people in the rural areas and the CSDs that

are distributed across the seven counties are more highly populated with less than 100,000 people. The City of London has the highest population density in the study area. Figure 3.1b categorized the population density at the DA level with each DA including 400 to 700 persons (Statistics Canada, 2017). The choropleth map showed a similar trend to the previous map with most areas being rural with population count less than 350, then areas with middle range count between 351 and 700, and lastly, the most highly populated areas which were located at several different pockets in and near the city of London.

The MLHU is the official health agency that provides information to improve quality of life for residents of the county. As indicated earlier, our choice of study area was made in collaboration with MLHU to promote breast cancer awareness in the region.

3.2.2 Breast cancer data

Breast cancer data was obtained from CCO who collaborated with health facilities to compile Ontario Cancer Registry, a database of patients diagnosed with cancer within the province. The database consists of incidents and mortality cases for all types of cancer but this study used only breast cancer cases among female patients between 2003 and 2013. A total of 97 cases with missing postal codes were removed and the remaining 7,771 valid postal codes were geocoded using Geocode tool in ArcMap 10.5 (ArcGIS software by Esri) and 2013 Multiple Enhanced Postal Code provided by DMTI Spatial Inc.

Table 3.1: Demographic data of counties and number of reported breast cancer incidents in 2003-2013

County	Area (km^2)	Number of CSD	Number of DA in 2011	Population in 2011	Breast cancer incidents (2003-2013)
Middlesex	3,318	12	713	439,151	3,381
Lambton	3,002	14		126,199	1,069
Oxford	2,040	8		105,719	737
Chatham-Kent	2,471	2		104,075	792
Elgin	1,881	8		87,461	637
Perth	2,218	6		75,112	572
Huron	3,400	9		59,100	583
Total	18,329	59		996,817	7,771

The detailed information regarding Middlesex County and the extended area with six counties that surround it are listed in Table 3.1. There is an increasing trend of breast cancer incidence in both Middlesex County and the surrounding counties (Figure 3.2).

Each recorded breast cancer case contains the birth year, the diagnosis year, and the residential postal code. Age, which was obtained by subtracting the diagnosis year from the birth year, is an important element of this study to measure breast cancer prevalence because age is shown to be highly correlated with the number of cancer incidents in our data (see Figure 3.3). Also, previous studies have reported that women in the age group of 50-74 have a higher risk of developing breast cancer than other age groups (Gail et al., 1989; Kelsey, 1993; McPherson et al., 2000). Hence, areas with a high population count of older women will likely have a larger number of breast cancer incidents; although, this does not translate to having a higher risk of the disease. Given that the number of cancer incidents alone may not explain the underlying distribution, this study factored in population counts for each age group to calculate age-adjusted cancer rates for each area. Comparing these rates can help to understand the existing patterns to potentially identify areas with high risk of breast cancer. Age-adjusted rates (AAR) are widely used in cancer studies to remove age bias in prevalence measurements.

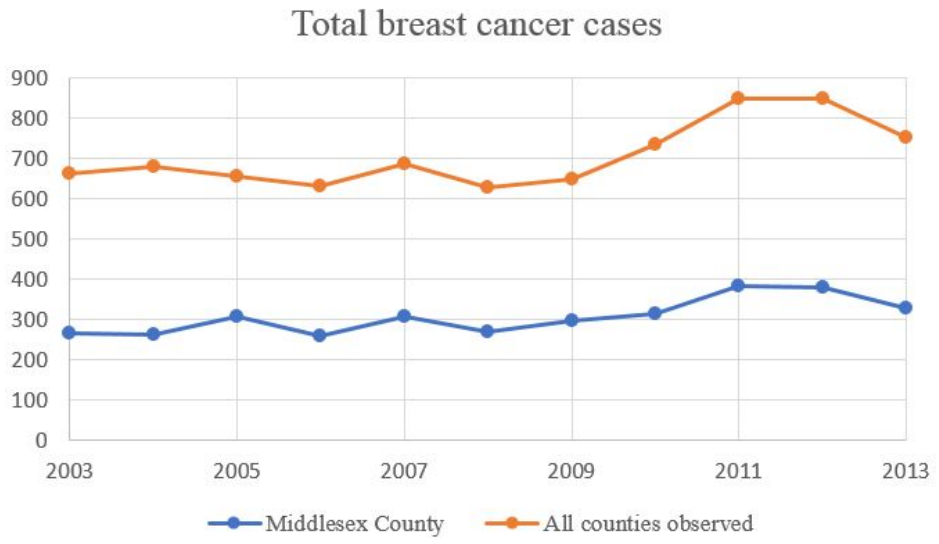

Figure 3.2: Breast cancer cases per year in Middlesex County (top line) and in all counties (bottom line).

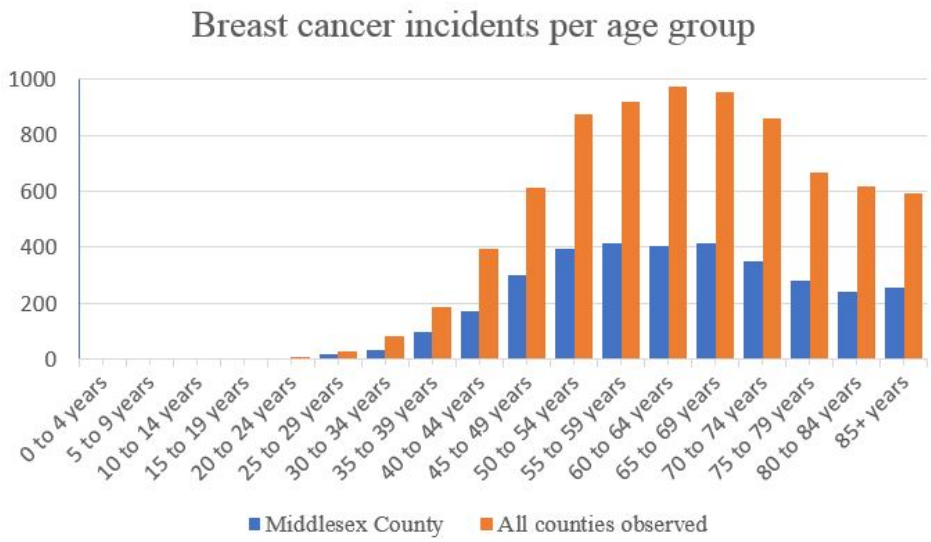

Figure 3.3: Breast cancer cases per age group in Middlesex County (shorter bars) in comparison to the number of cases within the extended area (longer bars). Both groups show a similar correlation between age increase and the number of cases.

3.2.3 Socioeconomic factors

The census program in Canada is conducted in two parts. Firstly, all households in Canada were asked to fill out a short questionnaire (census); then secondly, one-third of them were asked to fill out a more detailed questionnaire called the National Household Survey (NHS). This portion of the population is selected with a cross-sectional method to cover all persons who live in Canada including people who live on Indian reserves and on other Indian settlements, permanent residents, refugees and holders of work and study permits along with their family members. The subjects of the more detailed questionnaire include a) aboriginal status, b) education, training and learning, c) ethnic diversity and immigration, d) families, household and housing, e) income, pensions, spending and wealth, f) labour, g) languages, h) population and demography, i) society and community. Data from the more detailed questionnaire was used for this study. For each geographical unit, we used income, education, employment status, ethnicity, and occupation type from census data as socioeconomic covariates in the analysis.

3.2.4 Analysis

This section describes the summary of data preparation and analytical processes in this study. Several sets of data from various resources were collected and explored for integrity before we executed a series of geographical and statistical analyses. After the data preparation phase, breast cancer cases were projected into a list of areas, both at the DA and CSD levels, with their cancer count, population count, and map coordinates for geographical reference. Analyses were performed to obtain breast cancer prevalence, detect clusters of high and low values, and explore socioeconomic factors that could potentially be correlated with the increased risk of breast cancer in the study area. Figure 3.5 provides the flow of the processes that were implemented for the analysis.

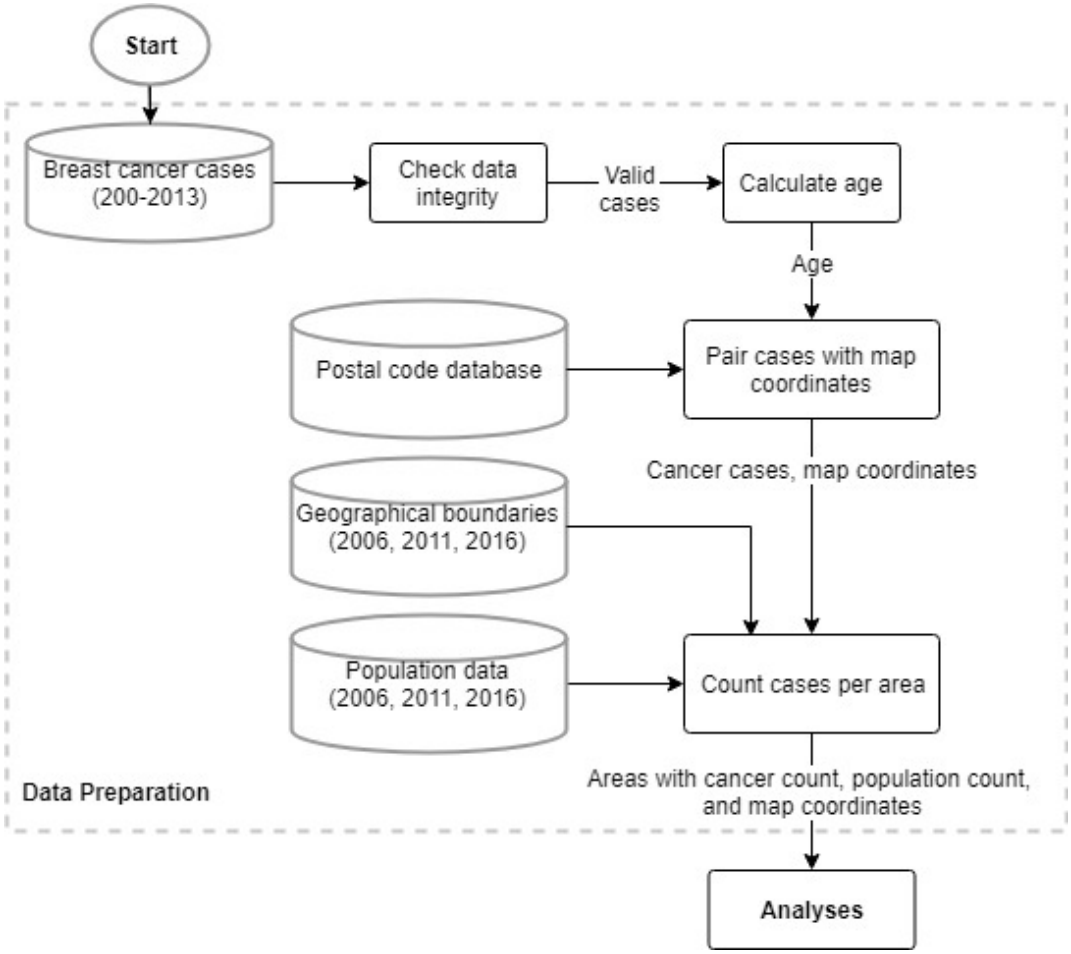

Figure 3.4: Flowchart of data preparation processes before the analysis

Geographical patterns of particular areas were influenced by factors including scale, thresholds, methodological issues, and the characteristic of the phenomena of interest. Objects in an area can be defined as following certain patterns if they form a line, delineate the contour of another set of objects, or are clustered within a geographical space. A cluster is a group of similar objects that are close to each other. In cancer studies, clusters are associated with areas that have a relatively higher number of cancer occurrences compared to other areas over a period of time (CCS, 2018). For instance, a map of breast cancer incident locations will be inadequate to explain the prevalence and burden over time. Hence, a spatial and temporal analysis is needed to detect any clusters that may be statistically significant. Various methods have been used to detect disease clusters, including analysis of point patterns, average nearest neighbour method, hotspot analysis, spatial autocorrelation, and spatial scan statistics. Anselin (2005) recommended the combination of the last two methods to compare the results for consistency and reliability.

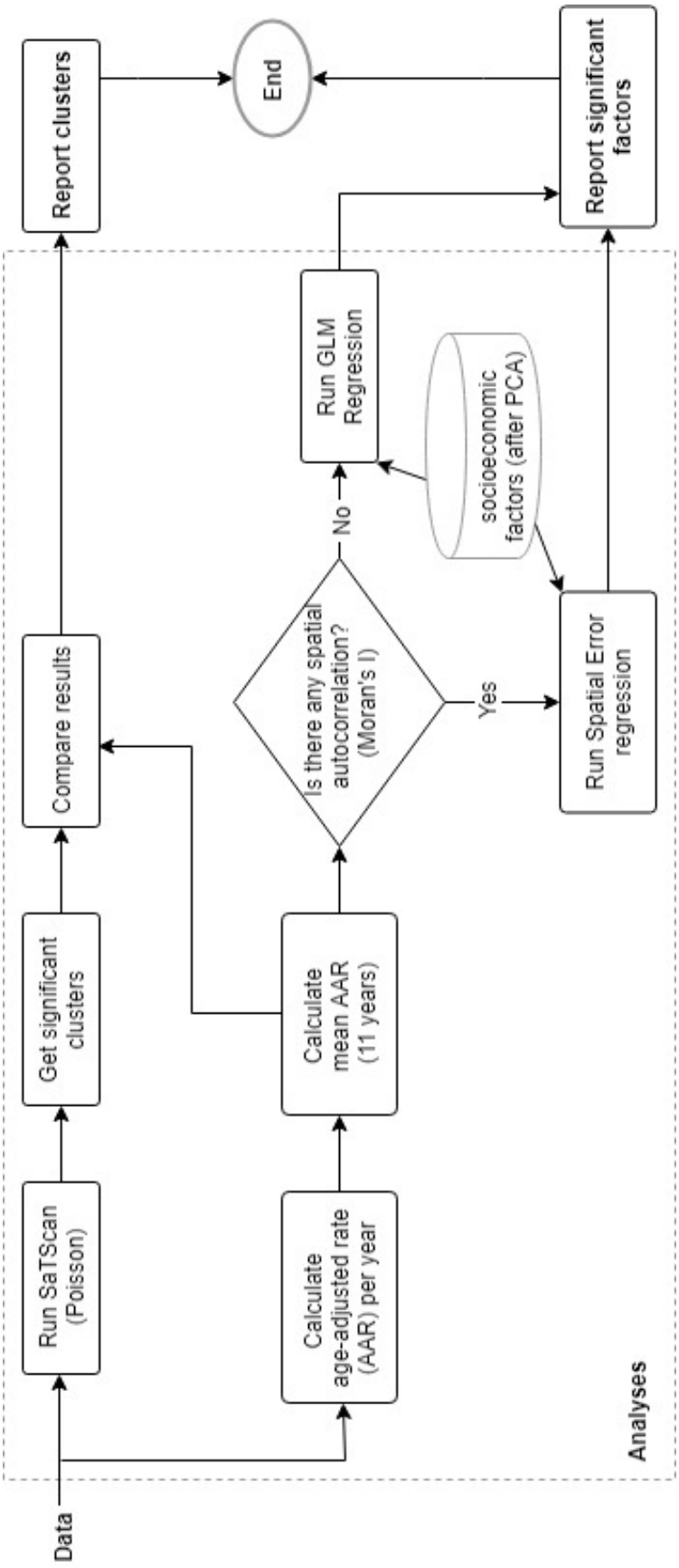

Figure 3.5: Flowchart of our analyses to detect clusters of breast cancer and to explore the correlation between breast cancer prevalence and socioeconomic factors

Age-Adjusted Rate (AAR)

Given the number of breast cancer cases in an area at a certain time and the total number of women living in that area, a crude rate can be calculated. If the population has a large number of women in higher age groups, then the crude rate may be inflated and potentially lead to a false conclusion that the area has a high breast cancer risk. For instance, retirement homes tend to have high counts of disease diagnosis and mortality rates because these conditions generally affect older people more than their younger counterparts. Such age confounding effects can be removed by factoring in the proportion of population within each age group as weights when generating disease rates, in this case, breast cancer, that are comparable across areas, regardless of the population age formations.

For each area and year, the number of cancer cases and the total population were distributed into 18 age groups (0-4 years, 5-9 years, and so on until 85+ years, see Table 3.2) and crude rates were calculated with the following formula:

$$CrudeRate = \frac{count}{population} * 100,000$$

The constant value of 100,000 is widely used to describe disease rates in epidemiology studies and the rate represents the number of people affected by the disease per 100,000 persons. AAR was calculated using the population proportion defined in the 2011 Canadian Standard Population. The weight for each age group was calculated as a proportion of the standard population for that age group over the total standard population, with the total weights of all age groups adding up to one. Lastly, each weight was multiplied by the corresponding crude rate to get the interim rate for the group and the sum of all interim rates generates the AAR. As an example, Table 3.2 displays the calculation with the resulting AAR of 252.63, which is significantly lower than the crude rate of 339.83 for the same data when the age effect has been removed.

$$CrudeRate = \frac{1,316}{387,251} * 100,000 = 339.83$$

Table 3.2: Example of AAR calculation using 2011 Canada Standard Population

Age group	Case count	Population	Crude rate	Standard Population	Weight	Rate
0-4 years	0	5,405	-	1,899,064	0.0553	-
5-9 years	0	7,077	-	1,810,433	0.0527	-
10-14 years	1	9,386	10.65	1,918,164	0.0559	0.60
15-19 years	5	11,864	42.14	2,238,952	0.0652	2.75
20-24 years	20	30,132	66.37	2,354,354	0.0686	4.55
25-29 years	25	21,612	115.68	2,369,841	0.0690	7.98
30-34 years	35	23,335	149.99	2,327,955	0.0678	10.17
35-39 years	37	23,836	155.23	2,273,087	0.0662	10.27
40-44 years	50	15,565	321.23	2,385,918	0.0695	22.32
45-49 years	61	32,726	186.40	2,719,909	0.0792	14.76
50-54 years	87	40,513	214.75	2,691,260	0.0784	16.83
55-59 years	118	38,946	302.98	2,353,090	0.0685	20.76
60-64 years	215	32,926	658.98	2,050,443	0.0597	39.34
65-69 years	225	26,552	847.39	1,532,940	0.0446	37.82
70-74 years	155	22,490	689.20	1,153,822	0.0336	23.15
75-79 years	148	24,330	608.30	919,338	0.0268	16.28
80-84 years	79	11,453	689.78	701,140	0.0204	14.08
85+ years	55	9,403	584.92	643,070	0.0187	10.95
Total	1,316	387,251	-	34,342,780	1	252.63

This study conducted spatial and temporal analysis at the DA and CSD levels, therefore AAR was also calculated at both levels. For each area, population data was retrieved from the 2011 Canadian Census available at Statistics Canada. CSD of Oneida 41 and Walpole Island did not have population data so they were excluded from the analysis. An AAR represented the rate for one area in one year, therefore the process was iterated for all areas for the 11 years study period. Two applications were written in Visual Basic for Application (VBA) for Microsoft Access to automatically generate the AAR values at the DA and CSD levels. Both applications used the same algorithms but were implemented using different datasets as described in Table 3.3.

Table 3.3: Comparison between analysis conducted at DA and CSD levels

Description	DA	CSD
Are areas stable during study period?	No	Yes
Number of areas	671 (2006)	59
Calculation of mean AAR	Raster processing to 2011 census	Calculate mean

A geographical area may change over time for administrative and electoral purposes and as the population grows or shrinks, areas may have a different number of households and persons resulting in areas being merged or split into different area boundaries. Furthermore, the Canadian Census is conducted every 5 years, thus the study period for this research intersected with three census times: 2006, 2011, and 2016.

CSD boundaries were stable throughout the study period so breast cancer prevalence at this level was retrieved by calculating the mean of AAR from all years. On the other hand, there were significant border changes during the observed years at the DA level and the differences affected AAR calculations. The area boundaries and population data were analyzed for all the earlier years and up to the particular census year. For example, the 2006 census data was used to compute breast cancer AAR between 2003 and 2006, and so on.

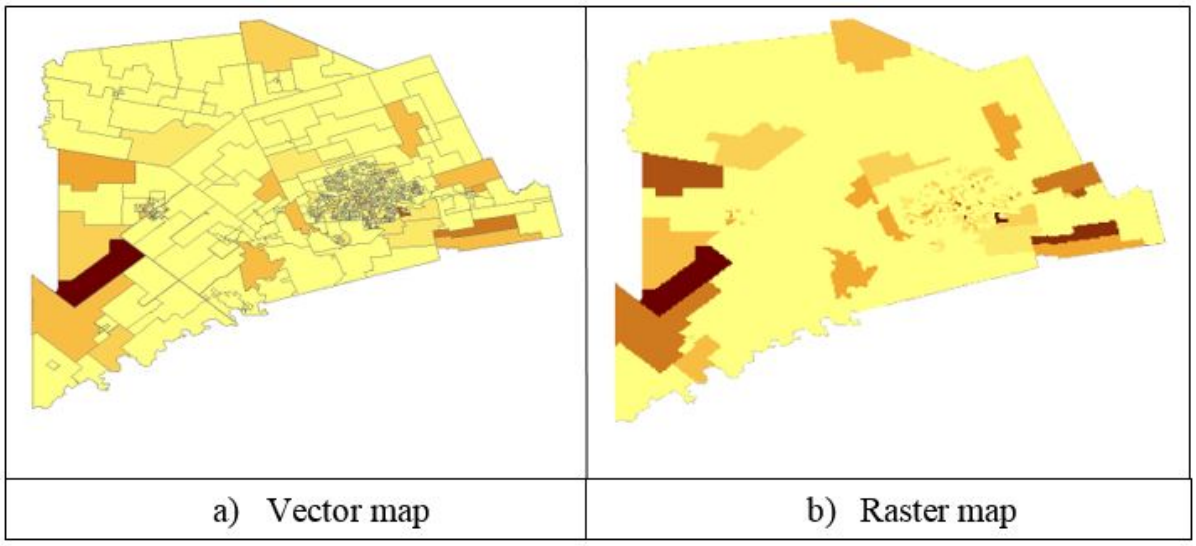

Figure 3.6: Transformation from a vector map of AAR to its raster form with mean AAR as the variable of interest

The study used raster analysis to accommodate different area boundaries. A map of AAR for each year was transformed into an image with each pixel representing an observed location containing AAR value. Figure 3.6a showed a map of AAR values in 2003 projected to 2006 census polygons showing various magnitudes of AAR values across the county. This map was transformed to a raster image (Figure 3.6b) using the 'Feature to Raster' tool in ArcMap

with each pixel representing the AAR value and the map symbology was set up to mimic the previous map for comparison purposes. Once this was done, the boundaries no longer existed in the raster image and each pixel stood on its own holding the AAR value for each location the pixel represented. The raster transformation was executed for all maps of AAR between 2003 and 2013. Since the pixels are not tied to specific boundaries anymore, the AAR average was calculated at the pixel level across 11 raster maps using the 'Raster Calculator' tool in ArcMap.

Spatial analysis for this study was conducted at the area level instead of the pixel level, hence the average AAR value per pixel was projected back into area boundaries. The ideal boundary map for this analysis was the maximum census year for the data, in this case, 2016, but with the study period stretching between 2003 and 2013, the use of the 2016 census would mean the data had a 3-year discrepancy with the maximum observation year and may distort the results. Therefore, the data used for the spatial statistical analysis was the census that was considered to best represent the population and socioeconomic factors during the study period, that is the 2011 census data.

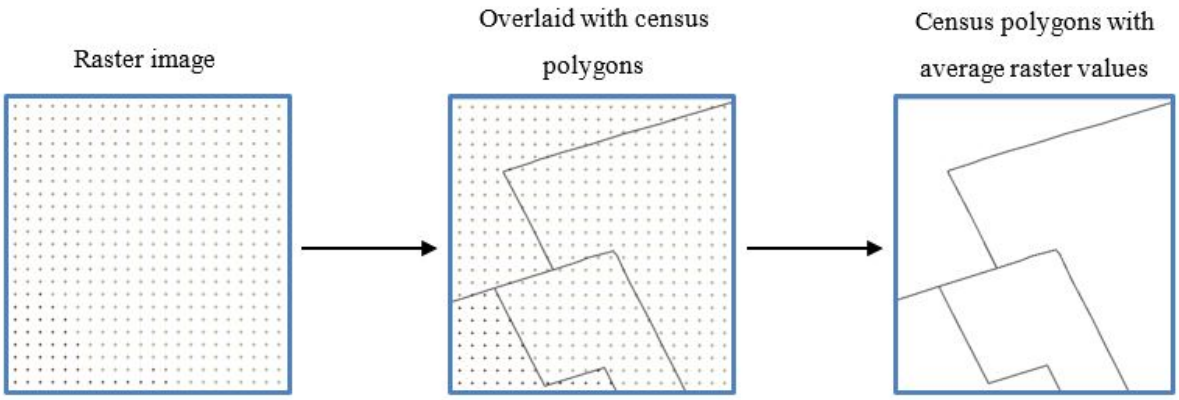

Figure 3.7: Spatial join process to amalgamate pixels in calculating mean AAR

Each pixel has the value of average AAR and they are assigned to the 2011 DA boundaries as illustrated in Figure 3.7. Such values depict breast cancer prevalence at the DA level.

Spatial scan statistic

The spatial scan statistic is a method used to test whether point patterns within space and/or time are purely random or follow particular patterns of clustering (Kulldorff, 1997). This statistic was used in our analysis because of its strength in detecting significant clusters while factoring in covariates that may create bias, such as age. The null hypothesis for this study is that breast cancer incidents happened randomly in geographic space and time proportional to the population at risk with the age bias removed. In other words, the risk of women getting breast cancer is the same everywhere in the study area. On the contrary, the alternate hypothesis holds that some underlying processes may trigger the pattern of breast cancer incidence that elevates the risk in a region.

Since breast cancer is a non-communicable disease, it is safe to assume that each case is independent. If each case happens without preceding conditions and is triggered by different unknown processes, then the occurrence is deemed to happen due to chance, or randomly. The characteristics of such cases fit the definition of a Poisson distribution, therefore this study performed cluster detection using SaTScan discrete Poisson analysis.

This method creates circular windows whose centres move around the area so that each window includes a different number of cancer cases. If a window contains the centre of a DA or CSD, then the whole area of that DA or CSD is included in the window. The centre of the window is positioned only at the center of DA or CSD with the radius from zero up to a maximum radius so that a window never includes more than half of the population of the area. A recommended guide for a window size is to include less than or equal to 50% of the population size because the cluster with a larger size would detect low rates outside the cluster rather than high rates inside the cluster (Kulldorff, 2015).

The result from this method is a list of windows that have a significantly higher likelihood of breast cancer risk compared to the areas outside the windows. The likelihood can be calculated with the formula below:

$$\left(\frac{n}{\mu}\right)^{n}\left(\frac{N-n}{N-\mu}\right)^{N-n} I(n > \mu)$$

where N = total number of incidents in the whole area

n = total number of incidents in the window

μ = age-adjusted expected number of incidents within the window

I = indicator function that is equal to 1 if the window has more incidents than expected, and 0 otherwise

The concept of the likelihood ratio is further illustrated in Figure 3.8. The box is considered an area with a total population of 2,000 people. Among these people, 100 are diagnosed with breast cancer, hypothetically depicted by the dots scattered inside the box. The likelihood of people being diagnosed with breast cancer in the whole population is as follows:

$$Likelihood_{population} = \frac{caseCount}{populationCount} = \frac{10}{2000} = 0.05$$

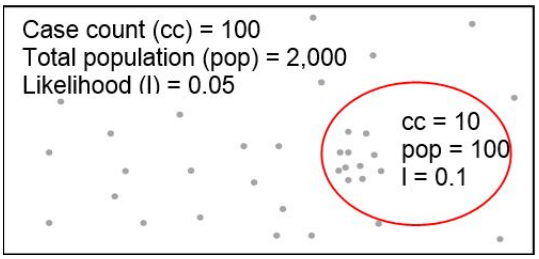

Figure 3.8: Illustration for likelihood ratio concept

The red circle inside the box illustrates a cluster detected by the software that has 10 cancer cases in it with the total population of 100 people inside the cluster area. The likelihood of people being diagnosed with breast cancer inside the cluster area is

$$Likelihood_{cluster} = \frac{caseCount}{populationCount} = \frac{10}{100} = 0.1$$

Lastly, the likelihood ratio for a specific cluster is calculated by dividing the likelihood inside

the cluster over the likelihood for the whole population.

$$LikelihoodRatio = \frac{Likelihood_{cluster}}{Likelihood_{population}} = \frac{0.1}{0.5} = 20$$

In this example, the likelihood ratio value reflects that the chance of being diagnosed with breast cancer is 20 times higher inside the cluster area compared to outside it. The method gradually scans windows across space and keeps a record of the observed and expected cases within the window. It calculates the likelihood ratio each time for the whole study area. Each window is a candidate for a cluster. Those with the highest likelihood ratios are the most likely clusters. Although spatial scan statistics can detect clusters in the shape of circles or ellipses, this study only used circles for simplicity.

Spatial scan statistics have been reported to be sensitive to the window size that is used for analysis (Boscoe et al., 2003; Ozdenerol et al., 2005; J. Chen et al., 2008). A large window size can result in failure to detect smaller clusters, on the other hand, a small window size may potentially lose large clusters that are significant. Yet within the context of breast cancer cluster detection, there is no specific maximum window size knowledge base to follow because the causes of breast cancer are still inconclusive. This study addressed this sensitivity by running the analysis with different window sizes and examining the results for consistency. We used SaTScan version 9.4.4 software written by Kulldorff (1997).

After the likelihood ratios for each cluster were calculated and sorted, a Monte Carlo simulation was performed to statistically test whether the ranking for our data happened due to chance at 95% confidence level. Windows with high likelihood ratios that were statistically significant were identified as clusters with high values and those with low likelihood ratios as clusters with low values. The analysis was run to detect space-time clusters, so the scanning windows include both space and time periods that are represented by cylindrical windows with the time as their height.

The results from SaTScan include a text output file with clusters information and a geographical output file (*shapefile*) that allows the clusters to be projected onto a map. All detected

clusters with high and low values were reported but we were only interested in those that were statistically significant. A process flow was built in ArcMap using the Model Builder tool to filter the results as shown in Figure 3.9. The results from this model allow the significant high and low clusters to be visualized and interpreted.

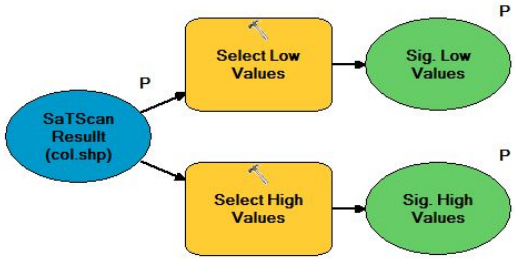

Figure 3.9: A process flow built in ArcMap to show significant high and low clusters

Spatial autocorrelation

Departing from Tobler's (1970) first law of geography that stated, *"everything is related to everything else, but near things are more related than distant things,"* the concept of spatial autocorrelation or spatial dependency measures how objects are similar to other objects located close to them (Anselin, 2005). The main distinction between non-spatial and spatial data is that the latter has inherent information about the locations of the data. In order to properly conduct spatial analysis, it is necessary to have additional information about the locations that indicate how closely the objects are situated in relation to other objects to which Tobler's law can be applied. This attribute is called the neighbour structure or spatial weights.

The choice of spatial weights is crucial when running such spatial dependency analysis. Spatial weights with distance band were considered less ideal for this study because the distances between DAs in the urban areas were far shorter than those in the rural areas. This study used the Queen's contiguity to create spatial weights in order to represent the adjacency of objects that takes all objects surrounding another object as its neighbours. The relatively stable number of neighbours for each area that follow a normal distribution was another reason for

the spatial weights selection.

The variable of interest for this method is the average value of AAR of breast cancer during the study period. A spatial autocorrelation tool helps to identify whether an area with a high rate was located close to areas that also have high rates, or whether an area with a high rate was surrounded by low rates and *vice versa*, or whether there were regions with uniformly similar (high or low) rates all around, or if there was any trend of rates at all.

A spatial autocorrelation is mostly measured using Geary Index and Moran's Index (Goodchild, 1986). Even though both indexes are robust for most applications, the latter provides a more intuitive result for interpretation with its positive or negative correlation. Within n number of areas the values to observe, the Morans I statistic was calculated as follows (Anselin, 1996):

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

For every two locations in the study area i and j, the attribute values of those locations were examined relative to their global mean value x and multiplied by w_{ij} which represents the value of spatial proximity between the two locations. The resulting index can range from -1 to +1. The former value indicates the lack of spatial dependency within the areas with high values surrounded by low values, or vice versa. This pattern is similar to a checkerboard pattern. The latter value represents areas with similar values that are grouped together to form clusters (high values surrounded by high values, and low values surrounded by low values). And lastly, the index value of zero can be interpreted as an absence of spatial autocorrelation for the variable of interest in the study area.

The results from cluster detection methods are dependent on the study area size, and as illustrated in Figure 3.8, the cases located in the circle are close to each other and they form a cluster because the distances between cases were much smaller when compared to the distances between cases in the whole area. However, the distances between cases on the left side of the circle tend to be similar to each other, thus the visible clusters at a smaller scale (or larger area) became less obvious. Due to the difference in perspective reasoning, this study conducted the

spatial autocorrelation analysis both at the global and local levels. We use the software *Geoda* v1.4.3 (TM) to perform spatial autocorrelation analysis.

Principal component analysis

Real world data can be sophisticated and the pursuit of modelling complex data requires careful attention. It is ideal to focus on the most important variables in the data. Principal Component Analysis (PCA) may help to determine variables that are more important than others by observing their variances. A variance is a measure of variability of a predictor (e.g.: total income, number of people with certain ethnicity). When we measure a predictor against a variable of interest (e.g.: breast cancer prevalence), a covariance is useful to determine the extent to which corresponding elements from the two variables move in the same direction. A positive covariance represents a positive relationship between the first and second variable, or in other words, as the value of the first variable increases, the other one does as well. On the contrary, a negative covariance describes the exact opposite relationship. When the two variables are not related, then the covariance equals zero.

In a model with many variables being considered, it is important to understand the covariance between variables to remove the possibility of statistical errors, both type 1 and type 2. A type 1 error is when we find a ”false positive” or a rejection of a true null hypothesis, and a type 2 error is failing to reject a false null hypothesis, also called a ”false negative” finding.

When data of a variable has a similar trend with another variable, one of them can be removed from the model. For example, human body weight and height generally increase with age. The similar trends for these two variables have less variability. In other words, a model that uses only one of these variables would be more compact without losing the variability of the data. PCA can help to measure the variability of data and reduce the number of variables used in the model to better represent the data and reduce complexity. In this analysis, PCA was utilized to examine the trends of various socioeconomic factors, retrieved from the NHS, and remove variables that show similar trends. For instance, education predictors may have a

similar trend of values with income variables because the fact that people with higher education tend to have a higher income is a general notion.

Multivariate analysis

Following the identification of cluster locations, socioeconomic factors were brought in to our analysis to explore common exposure patterns within the clusters. Census data in 2006 and 2011 were processed using their corresponding years of administrative boundaries and mean AAR.

We used multivariate analysis to examine the relationship between the dependent variable with other covariates. In fitting a model to a dataset, a set of weights was assigned to control for possible underlying statistical characteristics. Since larger sample sizes generate more reliable results (Costello & Osborne, 2005), we assumed that study areas with larger populations would yield more reliable breast cancer prevalences. To put this into practice, a general linear model (GLM) was used in which population counts were set as weights. The model was written in *R* and it is included in Appendix A. The software *R* is an open-source statistical and computing application developed at Bell Laboratories in New Jersey, United States.

Including a geographical aspect to a model is beneficial when we believe that the location of objects may affect the relationship of variables. Within the context of breast cancer, our analysis included geographical or spatial factors. If the spatial autocorrelation analysis provided evidence of geographical dependence in an area that influenced the values, then the bias of location can be removed by running a regression analysis that factors in a spatial continuity component. Anselin (2005) suggested a spatial regression decision process to determine the most appropriate model for spatial data and based on this suggestion, the spatial error regression model was deemed to be the best fit. Spatial error regression model was run with *Geoda* software and the results were interpreted within their assumptions and limitations.

3.3 Results

3.3.1 Clusters

SaTScan analysis was run at a series of maximum window sizes including 1%, 5%, 10%, 20%, 30%, 40% and 50% of the population at risk and the results are shown on the maps in Figure 3.10. Windows with significant maximum likelihood values were identified as clusters. The clusters are ranked in the order of their maximum likelihood values. With the 'Intersect' tool in ArcMap, the DAs that are located inside the clusters were marked as areas with significant risk of breast cancer, both high and low. The size of clusters ranges from zero to almost 15 kilometres in radius (see Table 3.5). A cluster with a radius of zero is represented with a dot on the map. This particular cluster is interpreted as a single DA with a significantly high or low maximum likelihood value.

When the maximum window size was set to 1%, SaTScan identified nine circles across the county with 22 DAs marked as clusters (see Figure 3.10) and when the window size was increased to 5%, SaTScan reported only seven circles with less amount of DAs (nDA=19). Similarly, the number of circles continued to decrease when the window size was increased to 10%. We began to see a saturation at the number of clusters detected at the 15% and 20% window sizes. Thus far, all of the DAs that were marked as clusters were located in the rural areas outside the City of London. Furthermore, when the window size was set to larger than 20%, the number of circles were unchanged but the number of included DAs began to exponentially increase, especially inside the circle located in the south-west of London (nDA>219) to fit the larger population size. This specific result could lead to an inaccurate detection of clusters, so the window size of 10% was chosen as the optimal parameter before saturation and to avoid potential errors.

SaTScan reported six significant clusters of high values and three clusters of low values in different periods of time and they are shown in Figure 3.11. This map also shows the prevalence of breast cancer measured with AAR for each DA using proportional symbology. Each

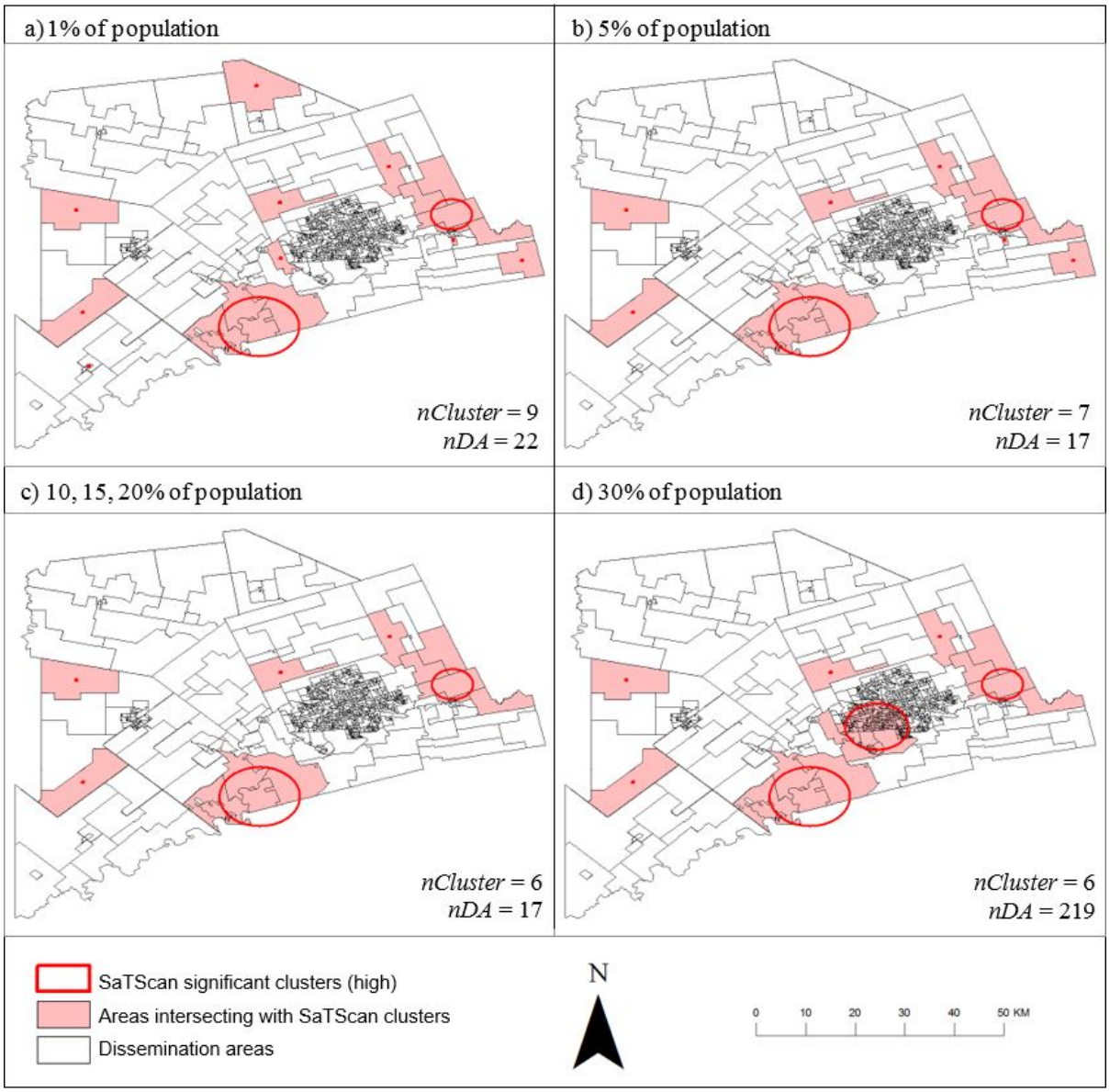

Figure 3.10: Maps displaying high-value clusters in Middlesex County with progression of an incremental population at risk percentages

circle depicts the ratio between AAR for the particular DA and the 2011 breast cancer national rate retrieved from the cancer burden report published by CCS (National AAR=130.8) (CCS, 2015). The larger the circle, the higher the risk of breast cancer in that particular DA compared to areas outside that DA, with the relative risk ratio to the national rate.

The use of proportional symbology and the chosen scale for Middlesex County (the top map in Figure 3.11) is helpful to visually scan the ratio distribution across the study area, but for small areas with many DAs, in this case, the City of London, the scale is too small to convey such distribution. A better representation of the city is shown at the bottom map using a larger map scale with the same proportional symbology size.

Table 3.4 shows DAs with the highest values of AAR and their comparison to the national rate. The column 'Ratio to NR' is calculated by dividing the AAR in the DA with the National Rate. The DAs in this list are marked as clusters with high values by SaTScan (clusters 1 to 6 in Figure 3.11). The analysis also revealed clusters of low values (clusters 7, 8, and 9) in the county. Apart from the area where clusters 6, 7, and 8 intersect, the areas marked as clusters of low values have low rate ratios as well, which confirms that the two methods yield similar results in measuring breast cancer prevalence. Both methods concluded that clusters of breast cancer tend to be located in the western and eastern fringes of the county.

Table 3.4: List of AAR at the DA level

AAR Category	DAUID	AAR	Ratio to NR	Cluster No.
9-10	35390825	1,371.10	10.48	1
	35390797	1,232.33	9.42	5
7-8	35390862	1,028.30	7.86	2
	35390883	981.10	7.50	4
	35390877	876.56	6.70	-
5-6	35390865	841.55	6.43	2
	35390751	727.11	5.56	3
	35390712	695.26	5.32	-
	35390723	664.65	5.08	6

National rate in 2011 = 130.8

For each identified cluster, SaTScan reported the radius, time frame, probability value, number of cases observed and expected, relative risk and likelihood ratio (Table 3.5). Clusters

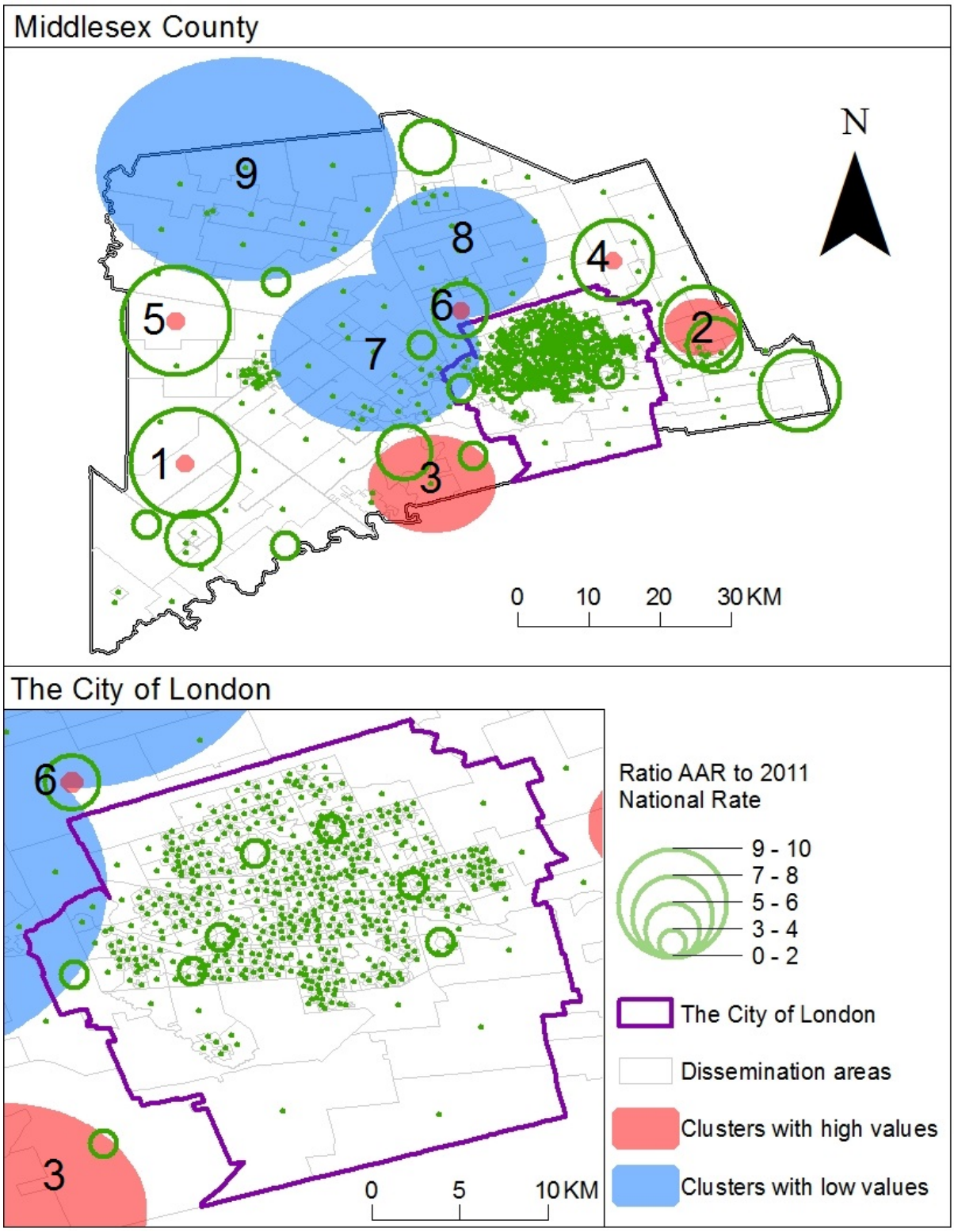

Figure 3.11: Clusters of breast cancer in Middlesex County identified by SaTScan overlaid on a map of proportionally symbolized ratio of age-adjusted rate to the 2011 national rate at the Dissemination Area level

1, 4, and 6 are persistently high between 2009 and 2013 and the rest of the clusters were reported at different time spans that vary across the study period.

Table 3.5: Details of significant clusters at the DA level

Type	No	Radius (km)	Start	End	P-value	Observed	Expected	RR*	LLR**
High	1	0	2009	2013	0.000000	24	1.55	15.54	43.31
	2	2.96	2003	2007	0.000000	33	3.90	8.52	41.46
	3	5.78	2010	2013	0.000000	25	2.43	10.37	35.80
	4	0	2009	2013	0.000000	19	1.82	10.47	27.40
	5	0	2005	2009	0.000007	16	1.44	11.19	24.04
	6	0	2009	2013	0.002872	18	3.04	5.96	17.11
Low	7	9.97	2008	2012	0.000000	7	53.98	0.13	33.01
	8	8.19	2004	2008	0.000028	0	22.33	0	22.40
	9	14.65	2004	2008	0.000108	0	20.78	0	20.84

*RR = Relative Risk, **LLR = Likelihood Ratio

A spatial and temporal analysis was also conducted to detect clusters at the CSD level for Middlesex County and its six land-locking counties to check the validity of our results. CSD boundaries were stable between 2003 and 2013, hence, breast cancer prevalence was retrieved by calculating the mean of AAR during the study period. Population data and geographical administrative boundaries from the 2011 census were used to generate the AAR.

Similarly to the map at the DA level, Figure 3.12 shows the locations of SaTScan clusters overlaid on a map of proportionally symbolized ratio between mean AAR per CSD and the 2011 breast cancer national rate, shown in Table 3.6. The areas with the highest prevalence are reported at four locations including Dawn-Euphemia CSD in Lambton County, Adelaide-Metcalfe CSD in Middlesex County, Howick CSD in Huron County, and Thames Centre CSD in Middlesex County (clusters 1 to 4). The map provides evidence that the results reported by the two methods are consistent with each other, in regards that the high-value clusters are located in the vicinity of CSDs with high rate ratios (larger circle) and similarly, the clusters with low values intersect with locations of CSDs with low rate ratio. With the focus on Middlesex County, the tendency of high values clustering are located at the western and eastern fringes of the county (cluster 2 and 4 in Figure 3.12) and this result is consistent with the previous result at the DA level (clusters 1 and 5 on the western side and clusters 2 and 4 on the eastern side

in Figure 3.11). The consistency shows that the result for our study area is not influenced by edge-effect bias.

Table 3.6: List of AAR at the CSD level

AAR Category	CSDUID	AAR	Ratio to NR	Cluster No.
>3	3538007	664.44	5.10	1
	3539047	646.83	4.95	2
	3540046	417.99	3.20	3
2	3539027	223.41	1.71	4

National rate in 2011 = 130.8

Table 3.7: Details of SaTScan significant clusters at the CSD level

Type	No	Radius (km)	Start	End	P-value	Observed	Expected	RR*	LLR**
High	1	0	2009	2013	0.000000	46	7.23	6.39	46.43
	2	0	2009	2013	0.000000	44	9.53	4.64	32.90
	3	0	2004	2008	0.000000	42	10.99	3.84	25.35
	4	0	2007	2011	0.023000	82	46.26	1.78	11.28
Low	5	0	2009	2012	0.000001	0	23.23	0.00	23.27
	6	0	2009	2013	0.000002	0	21.65	0.00	21.68
	7	17.13	2006	2010	0.003222	19	51.28	0.37	13.48
	8	14.56	2009	2013	0.004501	23	56.84	0.40	13.10

*RR = Relative Risk, **LLR = Likelihood Ratio

3.3.2 Spatial autocorrelation

Spatial autocorrelation analysis was run at both global and local levels to observe spatial patterns of AAR among neighbouring areas for individual and cumulative years. The variation in the number of years included in the analysis was useful to check spatial patterns consistency.

Global Moran's Index

We explored patterns of AAR values for each year between 2003 and 2013 and the results are shown in a series of scatter plots in Figure 3.13. The x-axis represents the standard deviation of AAR and the y-axis shows the weighted average of AAR in its neighbouring areas. The slope coefficients in the graph represent the Global Moran's Indexes, which range from -0.13 up to 0.05. The coefficients formed trend lines that are almost parallel to the zero axes indicating a

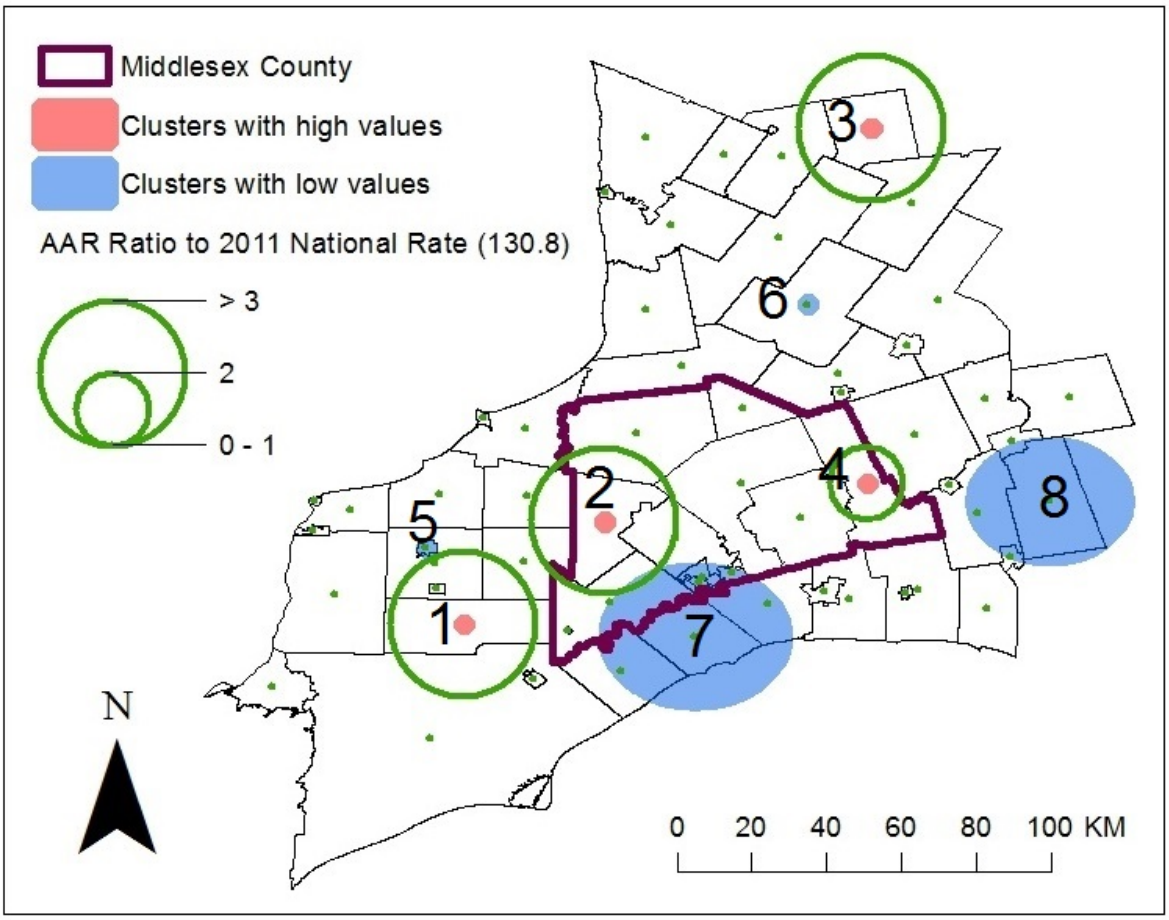

Figure 3.12: Clusters of breast cancer in Middlesex County and its surrounding counties overlaid on a map of proportionally symbolized ratio of age-adjusted rate to the 2011 national rate at the Census Sub-Division level

weak global spatial autocorrelation in any given year. Furthermore, none of the indexes were reported significant at the 95% confidence level by the Monte Carlo simulation tests. Conversely, the global analysis at the DA level for cumulative years reported a significant Global Moran's Index of 0.07 (p<0.01, 95% CI) indicating a weak positive spatial autocorrelation in the county.

Local Indicators of Spatial Analysis (LISA)

AAR values were also explored using the Local Indicators of Spatial Association (LISA) tool to detect local spatial autocorrelation per year. Each map in Figure 3.14 shows areas with similar values that cluster together having high or low values (High-High and Low-Low) and areas with contrasting values (High-Low and Low-High). Even though some areas are marked a few times as clusters of low values in the western part of the county and some other areas are marked the opposite, these maps convey inconsistent patterns of breast cancer across the study area.

The LISA result for cumulative years is shown in Figure 3.15. There is a clear sign of spatial heterogeneity in the study area with low-low values situated in the areas around the centre and the north-west corner and high-high values located in different pockets of the county, notably on the western and eastern parts. The reported areas returned by this method are relatively consistent with clusters detected with SaTScan in Figure 3.11. The significant existence of spatial autocorrelation in the study area determines multivariate analysis methods used to explore the link between breast cancer and socioeconomic factors.

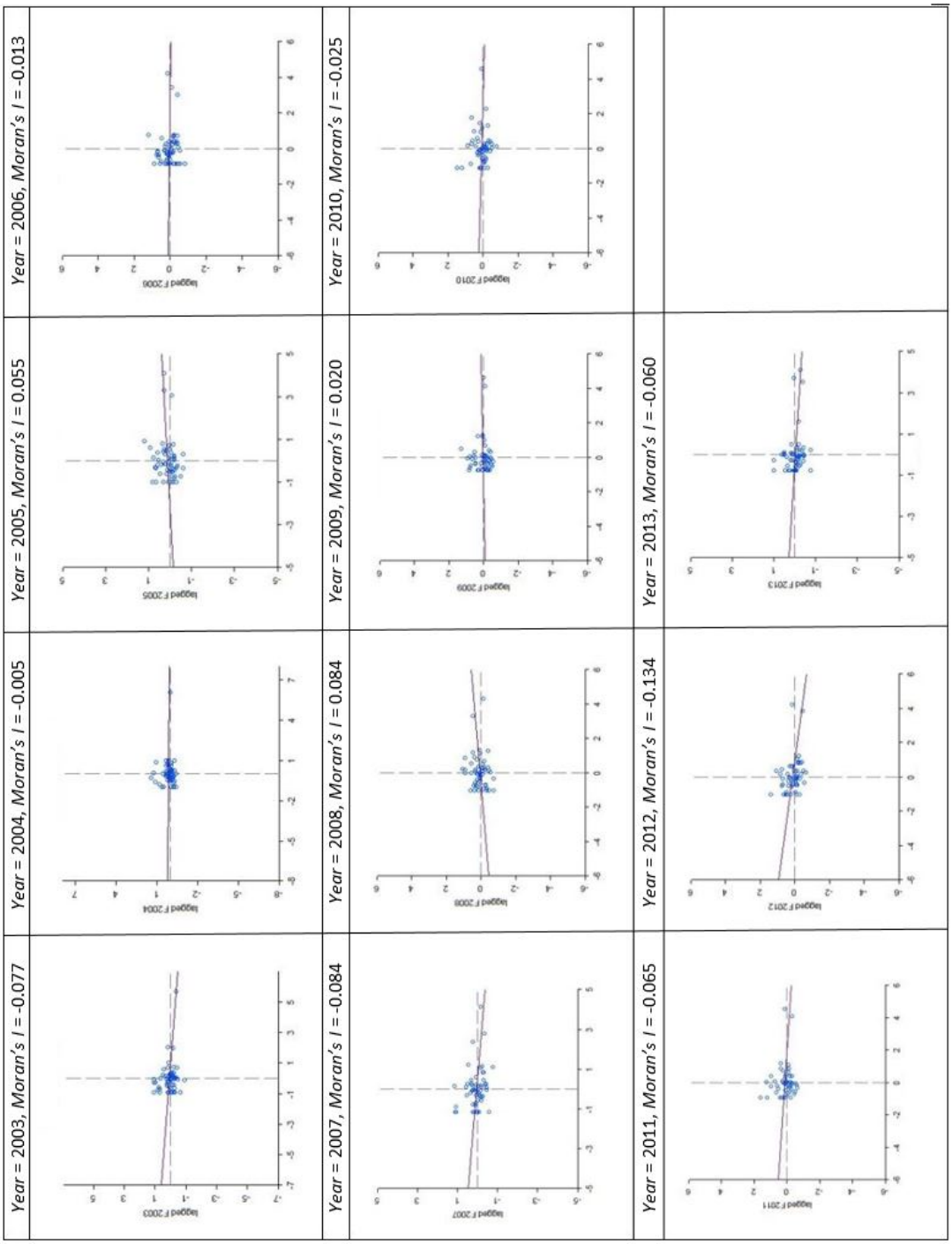

Figure 3.13: Scatter plots of spatial autocorrelation for each year between 2003 and 2013

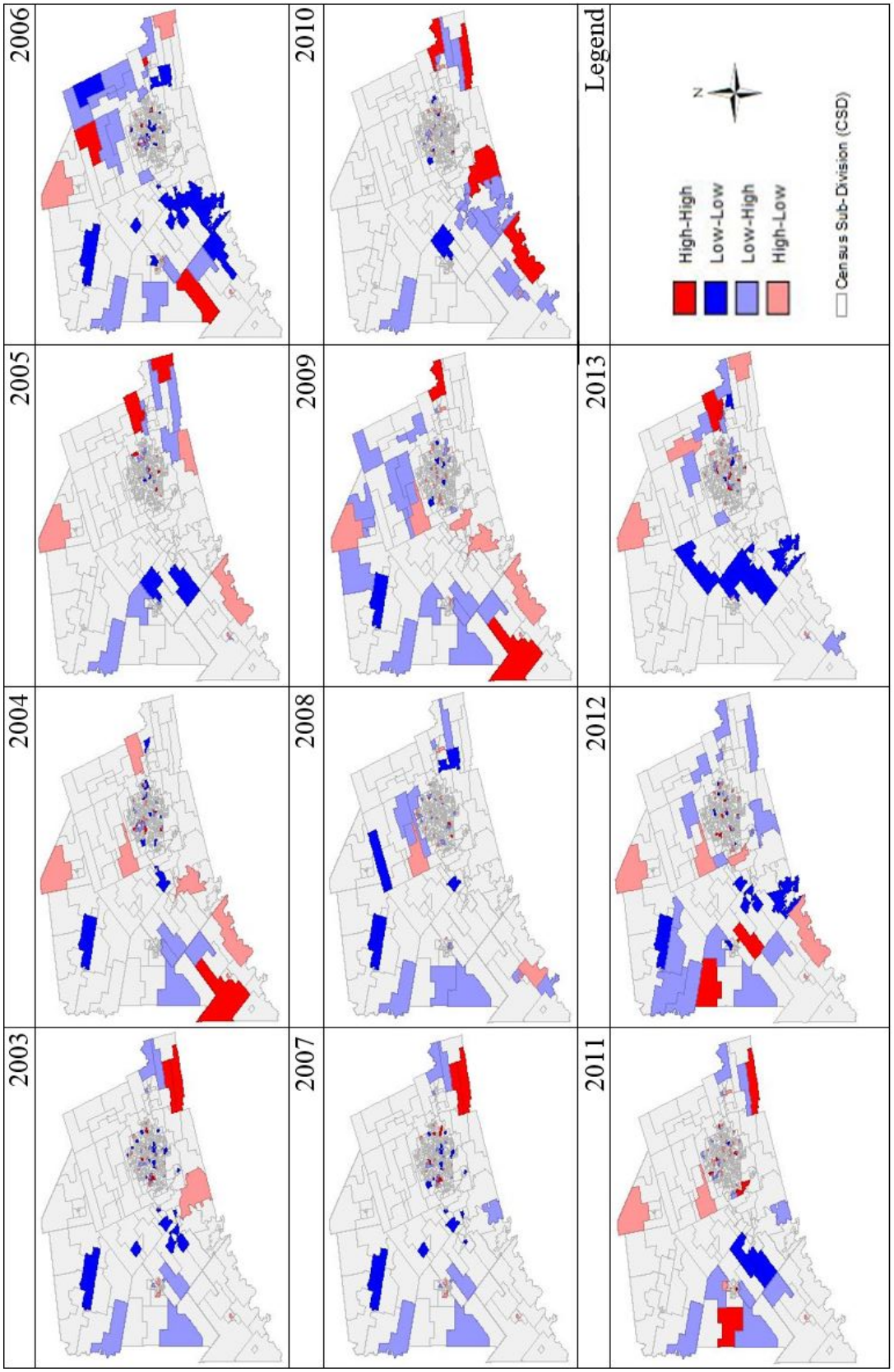

Figure 3.14: Patterns of local spatial autocorrelation at the DA level for each year between 2003 and 2013

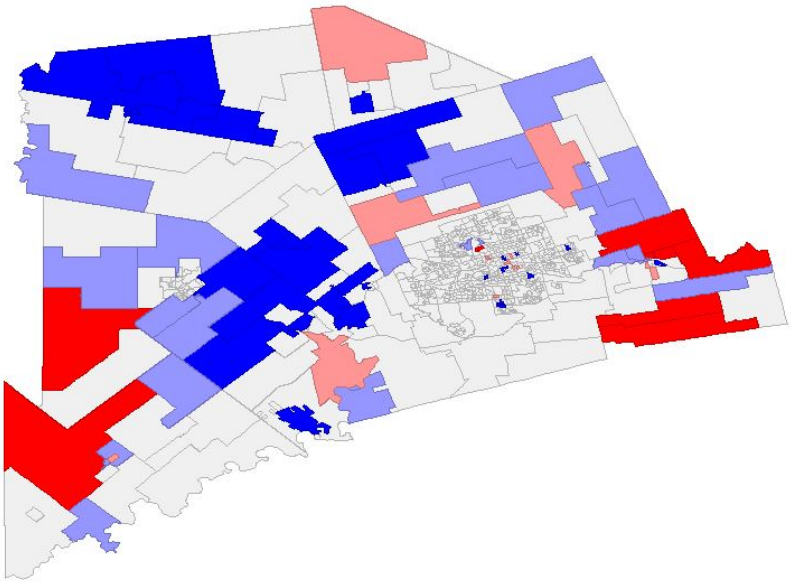

Figure 3.15: The average pattern of local spatial autocorrelation of breast cancer AAR between 2003 and 2013

3.3.3 Socioeconomic factors

We explored correlations between the identified breast cancer clusters and socioeconomic factors in the region. After careful selection of the variables available on the NHS based on other studies that have identified socioeconomic factors linked to breast cancer, we retrieved 31 independent variables from each of the 2006 and 2011 census data. The variables include various factors of income, education, employment, ethnicity, and occupation which were analyzed with multivariate analysis. The full list of variables and definitions are listed in Appendix A.

All variables were included in the GLM model for each census year at both DA and CSD levels using the corresponding population count as weights and the models came out insignificant. We then explored the possibility to reduce the model complexity by running PCA for each analysis. PCA allowed the models to be simplified by including fewer variables, reported as principal components, while the data still held reasonably high variance. The cumulative proportion of variability was chosen at 90% at the CSD level including 5 principal components for each census year. Meanwhile, at the DA level, the data variance was distributed more evenly

between the variables. The number of variables included in the model after PCA went down by about one third to reduce complexity while still maintaining the variance at 70%. The list of principal components for each census year and geographical unit is displayed in Table 3.8.

Table 3.8: List of variables with the highest variability in the data after PCA

Geographical Unit	Principal components	
	2006	2011
Census Sub-Division	Employment rate Average income Occupation in agriculture, forestry, fishing, and hunting Occupation in utilities Occupation in mining quarrying and oil and gas extraction	Employment rate Average income Occupation in agriculture, forestry, fishing, and hunting Occupation in utilities Occupation in mining quarrying and oil and gas extraction
Dissemination Area	Occupation in agriculture, forestry, fishing, and hunting Average income Employment rate Occupation in information and cultural industries Occupation in arts entertainment and recreation Aboriginal ethnicity Occupation in transportation and warehousing Occupation in construction Occupation in wholesale trade Occupation in real estate and rental and leasing	Occupation in agriculture, forestry, fishing, and hunting Employment rate Occupation in arts entertainment and recreation Average income Occupation in construction Aboriginal ethnicity Occupation in information and cultural industries Occupation in wholesale trade Occupation in transportation and warehousing Occupation in real estate and rental and leasing Total population with no certificate diploma or degree

Multivariate analysis was then conducted using principal components for the 2006 and 2011 census data at both the DA and CSD levels. The absence of spatial autocorrelation at the CSD level justified the use of GLM for our multivariate analysis with population count as weights. On the contrary, spatial autocorrelation was significantly detected at the DA level. The use of GLM for the data at this level would generate inaccurate results because of the inherent spatial bias in the data, therefore, spatial error regression was chosen instead. Table 3.9 displays the multivariate analysis results. The models consistently reported three factors that are significant including average income, employment rate, and the collective occupations in agriculture, forestry, fishing and hunting.

Table 3.9: Multivariate analysis results

Predictors	General Linear Model		Spatial Error Reg.	
	CSD		DA	
	2006	2011	2006	2011
Average income	0.00	-0.04 ***	0.02 ***	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Employment rate	0.87	2.95 ***	0.08	1.50 **
	(2.28)	(0.42)	(0.52)	(0.55)
Occupation in agriculture,	-0.03	-0.65 *	3.43 ***	7.37 ***
forestry, fishing and hunting	(0.09)	(0.27)	(0.96)	(2.18)

Significance: *** p<0.001, ** p<0.01, * p<0.05

The employment rate for women who lived in the study area was not correlated to the increased risk of breast cancer in 2006, but it was statistically significant in 2011 at both CSD and DA levels. At the CSD level, the result shows that employed women were almost three times more likely to develop breast cancer compared to women who were not employed. Meanwhile, at the DA level, the likelihood was also significant, but only by half as much.

The risk of breast cancer was reported to slightly increase with a small elevation of average income from the analysis at the DA level in 2006, but conversely, breast cancer incidence was correlated to a slight decrease of average income when the analysis was conducted at the CSD level using 2011 census data.

Furthermore, a significant link was found between breast cancer development and the risk factor that included occupation in agriculture, forestry, fishing and hunting. The analysis at the CSD level in 2006 yielded no significant result and the 2011 data shows that women who worked in this set of occupations were slightly less likely to develop breast cancer. A much stronger correlation is reported for this particular risk factor from the analysis at the DA level. In 2006, women in this set of occupations were three times more likely to develop breast cancer in 2006 and by 2011, the likelihood of developing breast cancer increased to more than seven times compared to other women who did not work in these fields.

3.4 Discussion

Cluster detection with SaTScan showed a sensitivity of outcome to the window size used in the analysis. The higher number of population count is allowed to be included in a cluster, the bigger the cluster size may become. In such case, the size of the detected cluster may not portray the real burden of breast cancer in the area, but the incorporation of a larger number of people may lead to an inaccurate detection of clusters. Small clusters may be concealed by a large cluster because the test statistic value for the large cluster may be significant due to the inclusion of the smaller significant clusters that are surrounded by insignificant ones. The progressive comparison of window sizes was necessary to achieve the optimum window size for our data.

There is no evidence of a strong pattern of breast cancer clustering in our study area at the CSD level, but the analysis at the DA level identified the locations of high-value clusters in rural areas. Residents of rural areas have been linked to having poorer health status than their urban counterparts (Pong et al., 2009) and they have also been reported to be less likely to seek preventive services compared to women who live in urban areas (Bryant & Mah, 1992). One possible explanation is that rural areas tend to have less accessible health services. The Canadian Partnership Against Cancer reported that only a smaller percentage of rural patients are able to access cancer treatment compared to their urban counterparts, especially those who live further away from the treatment facilities (*2014/2015 Annual Report: Progress in Action*, 2014). Consequently, women in rural areas might delay mammogram checks or wait longer for test results and these actions may lead to late detection and treatment, which then increase breast cancer mortality rate.

The results from this study reported the socioeconomic factors related to an increased breast cancer risk include income, employment rate, and occupations in agriculture, forestry, fishing and hunting.

The correlation between breast cancer and women's average income are sparingly weak but significant across our analyses at different geographical units and census years. Income and

employment rate may be treated as indicators of socioeconomic status(SES). Many studies have reported the link between SES and breast cancer development. At the individual level, breast cancer risk factors have been reported to include genetics, the age of first menstruation, parity, age, hormonal variations, family history (Lynch et al., 1989; Trichopoulos et al., 1983; Ye et al., 2002; Lanfranchi, 2015). A study conducted in Wisconsin controlled for these individual level risk factors and found that socioeconomic factors at the community level, including high level of education, urban life, and living in high SES community, also contributed to the increased risk of breast cancer (Robert et al., 2004). They reported community contextual effects may influence individual-level choices and behaviours regarding reproductive and lifestyle factors that may increase the odds for women to develop breast cancer in their lifetime. Higher SES women have been reported to have the tendency to have lower parity, later age for first full-term pregnancy, greater body weight, higher alcohol consumption, lower lactation, exogenous hormone use, and greater use of mammography screening (Kelsey & Bernstein, 1996; Katz et al., 2000; Calle et al., 1993)

Furthermore, women who work may be subjected to environmental exposures that exist at their workplaces, both for the purpose of the job (e.g.: the use of chemicals in a factory, or cleaning products used by women who work as cleaners, etc.) as well as the workplace environment contaminants that may trigger the development of breast cancer. There are many facets of employment that can be derived to explain its link to breast cancer, which makes this particular economic factor less conclusive in explaining breast cancer occurrence.

Among the socioeconomic factors explored in this analysis, our findings show the strongest correlation between an increased risk of breast cancer and occupations in agriculture, forestry, fishing and hunting. Even though there is a lack of evidence for the association between the last two occupations, fishing and hunting, and breast cancer, some studies have reported links between environmental aspects of the first two occupations and breast cancer development (Weiderpass et al., 2011; Fenga, 2016; X. Chen, 2017). For the occupation in agriculture, the clusters detected in this study are mostly located in rural areas and these are consistent with the

general locations of the lands used for agriculture. Farmers had been reported to experience elevated rates of cancer due to their exposure to a variety of substances including pesticides, engine exhausts, solvents, dust, and zoonotic microbes (Blair & Zahm, 1995). Furthermore, this county sits in the area of south-western Ontario, which is known to be an agricultural region because the climate is among the mildest in Canada. This finding is consistent with earlier work by Brophy et al. (2012) who reported that women who had worked in agriculture were more susceptible to developing breast cancer.

Solar radiation has been reported as an occupational exposure in forestry that may promote cancer development, although the results are inconsistent with breast cancer high prevalence. Meanwhile, studies that conclude an association between occupation in forestry argue that workers may be exposed to some carcinogens found in organic compounds when handling wood (Guénel & Villeneuve, 2006). The carcinogens were reported to have the same properties as an endocrine disruptor that has been linked to an increased risk of breast cancer.

Study Limitation

In census data, some areas were reported to have a population of zero. In reality, these areas likely have dwellers living in them, but the voluntary nature of census data collection might lead to an incomplete representation of the population count. When the analysis was run for these areas, the generation of AAR, that included a calculation of cancer count divided by the population count of zero, resulted in an unknown number of infinity and the rates could not be used for the analysis. The prevalence of breast cancer in these areas are not properly projected in our AAR measurement causing a distortion in the results. If the numbers of breast cancer incidents were high in these areas, then our analysis might have missed potentially significant clusters.

The results from this study need to be interpreted with caution considering uncertainties in the latency period during which the disease develops. The unrecorded history of migration may contribute to inaccuracy in the results. Some women might decide to live or retire in the

areas that are more accessible for them, some others might move to big cities for easy access to health care, and some might migrate for other significant reasons. Women could be exposed to environmental factors that would increase breast cancer risk before or during the migration, but the cancer registry only captures the information at the time of cancer diagnosis.

The response rate for the 2006 Canadian census was 93.8% and it went down to 68.6% in 2011 (Statistics Canada, 2018). Even though Statistics Canada had taken an adjustment approach to improving data quality, the estimates we retrieved from the NHS were released with caveats to inevitably include potential challenges of the variability of response rates at lower geographic levels, sampling error and non-response bias. Furthermore, the aggregation of health data in our study, from the postal code level to the DA or level, is susceptible to ecological fallacy and reasoning inaccuracy.

3.5 Conclusion

The study used both spatial autocorrelation and spatial scan statistics to identify breast cancer clusters in Middlesex County. The analyses were conducted within the county and with the inclusion of its surrounding counties to check the consistency of results and they were reasonably similar to each other. Areas identified as clusters of high values were reported at the western and eastern fringes in the county as shown in Figure 3.10 and Figure 3.12. Clusters of low values were also detected and they were located in the northern part of the county.

This research was conducted in collaboration with the Middlesex-London Health Unit and the results may contribute to local breast cancer programs and policy changes that are focused on raising awareness about the disease in the community. Early detection of breast cancer cases applied through the screening programs may lead to improved survival rates. Women who are currently eligible for breast cancer screenings are categorized into those at average risk (aged 50 to 74 years) and those at high risk (aged 30 to 69 years) (CCO, 2018). The former group consists of women without acute breast symptoms, nor a personal history

of breast cancer, nor current breast implants, and not having had a mammogram in the last 11 months. The high-risk group includes women with a genetic mutation, family history of breast cancer, and those who received radiation therapy on their chest. Based on the findings in this study, we recommend the health unit to specifically encourage women with low income, high unemployment rate, and occupation in agriculture, forestry, fishing, and hunting to also participate in the screening program.

References

2014/2015 Annual Report: Progress in Action (Tech. Rep.). (2014). Canadian Partnership Against Cancer. Retrieved from https://www.partnershipagainstcancer.ca/about-us/corporate-resources-publications/annual-reports/

Anselin, L. (1996). The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In *Spatial analytical perspectives in gis* (pp. 111–125). Taylor & Francis. Retrieved from http://dces.wisc.edu/wp-content/uploads/sites/30/2013/08/W4{_}Anselin1996.pdf

Anselin, L. (2005). Exploring Spatial Data with GeoDa: A Workbook. *Geography*, 244. doi: http://www.csiss.org/

Blair, A., & Zahm, S. H. (1995, nov). Agricultural exposures and cancer. *Environmental health perspectives*, *103 Suppl*(Suppl 8), 205–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8741784http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1518967

Boscoe, F. P., McLaughlin, C., Schymura, M. J., & Kielb, C. L. (2003, sep). Visualization of the spatial scan statistic using nested circles. *Health and Place*, *9*(3), 273–277. Retrieved from http://www.sciencedirect.com/science/article/pii/S1353829202000606 doi: 10.1016/S1353-8292(02)00060-6

Brophy, J. T., Keith, M. M., Gorey, K. M., Luginaah, I., Laukkanen, E., Hellyer, D., ... Gilbertson, M. (2006). Occupation and breast cancer: A Canadian case-control study. *Annals of the New York Academy of Sciences*, *1076*, 765–777. doi: 10.1196/annals.1371.019

Brophy, J. T., Keith, M. M., Watterson, A., Park, R., Gilbertson, M., Maticka-Tyndale, E., ... Luginaah, I. (2012). *Breast cancer risk in relation to occupations with exposure to carcinogens and endocrine disruptors: a Canadian casecontrol study* (Vol. 11) (No. 1). Toronto: National Network on Environments and Women's Health. Retrieved from http://ehjournal.biomedcentral.com/articles/10.1186/1476-069X-11-87 doi: 10.1186/1476-069X-11-87

Bryant, H., & Mah, Z. (1992, jul). Breast cancer screening attitudes and behaviors of rural and urban women. *Preventive Medicine*, *21*(4), 405–418. Retrieved from https://www.sciencedirect.com/science/article/pii/009174359290050R doi: 10.1016/0091-7435(92)90050-R

Calle, E. E., Flanders, W. D., Thun, M. J., & Martin, L. M. (1993, jan). Demographic predictors of mammography and Pap smear screening in US women. *American journal of public health*, *83*(1), 53–60. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8417607http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1694510

Calle, E. E., Frumkin, H., Henley, S. J., Savitz, D. A., & Thun, M. J. (2002, sep). Organochlorines and Breast Cancer Risk. *CA: A Cancer Journal for Clinicians*, *52*(5),

301–309. Retrieved from http://doi.wiley.com/10.3322/canjclin.52.5.301 doi: 10.3322/canjclin.52.5.301

Canadian Cancer Society. (2015). Canadian Cancer Statistics Special topic : Predictions of the future burden of cancer in Canada. *Public Health Agency of Canada*, 1–151. doi: CanadianCancerSociety

Canadian Cancer Society. (2018). *Cancer clusters.* Retrieved 2018-05-05, from http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/cancer-clusters/?region=on

Cancer Care Ontario. (2018). *Ontario Cancer Statistics 2018 Report.* Retrieved 2018-10-01, from https://www.cancercareontario.ca/en/statistical-reports/ontario-cancer-statistics-2018-report

Chen, J., Roth, R. E., Naito, A. T., Lengerich, E. J., & MacEachren, A. M. (2008, nov). Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. *International Journal of Health Geographics*, *7*(1), 57. Retrieved from http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-7-57 doi: 10.1186/1476-072X-7-57

Chen, X. (2017). A temporal analysis of the association between breast cancer and socioeconomic and environmental factors. *GeoJournal*, 1–18. doi: 10.1007/s10708-017-9824-5

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. - Practical Assessment, Research & Evaluation. *Practical Assessment Research & Evaluation*, *10*(7). Retrieved from https://www.pareonline.net/pdf/v10n7.pdf

Crouse, D. L., Goldberg, M. S., Ross, N. A., Chen, H., & Labrèche, F. (2010, nov). Postmenopausal breast cancer is associated with exposure to traffic-related air pollution in Montreal, Canada: A case-control study. *Environmental Health Perspectives*, *118*(11), 1578–1583. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20923746http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2974696 doi: 10.1289/ehp.1002221

Fenga, C. (2016). Occupational exposure and risk of breast cancer. *Biomedical Reports*, *4*, 282–292. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4774377/pdf/br-04-03-0282.pdf doi: 10.3892/br.2016.575

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, *81*(24), 1879–1886. doi: 10.1093/jnci/81.24.1879

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). *Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology* (Vol. 21; Tech. Rep. No. 1). Retrieved from https://www-jstor-org.proxy1.lib.uwo.ca/stable/pdf/622936.pdf

Goodchild, M. F. (1986). *Spatial Autocorrelation*. Geo Books.

Guénel, P., & Villeneuve, S. (2006). Occupational exposure to organic solvents and breast cancer in men and women: new results strengthen the hypotheses of environmental risk factors. *French Institute for Public Health Surveillance*. Retrieved from file:///C:/Users/Jenny/Downloads/plaquette{_}occupational{_}exposure{_}organic{_}solvents{_}breast{_}cancer{_}men{_}women{_}results{_}strengthen{_}hypotheses{_}environmental{_}risk{_}factors.pdf

He, T.-T., Zuo, A.-J., Wang, J.-G., & Zhao, P. (2017, may). Organochlorine pesticides accumulation and breast cancer: A hospital-based casecontrol study. *Tumor Biology*, *39*(5), 101042831769911. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28459199http://journals.sagepub.com/doi/10.1177/1010428317699114 doi: 10.1177/1010428317699114

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., & Starosławska, E. (2015, sep). Breast cancer risk factors. *Przeglad menopauzalny = Menopause review*, *14*(3), 196–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26528110http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4612558 doi: 10.5114/pm.2015.54346

Katz, S. J., Zemencuk, J. K., & Hofer, T. P. (2000, may). Breast cancer screening in the United States and Canada, 1994: socioeconomic gradients persist. *American journal of public health*, *90*(5), 799–803. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10800435http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1446215

Kelsey, J. L. (1993). Breast cancer epidemiology: summary and future directions. *Epidemiologic reviews*, *15*(1), 256–263. Retrieved from https://www.popline.org/node/328958

Kelsey, J. L., & Bernstein, L. (1996, jan). Epidemiology and Prevention of Breast Cancer. *Annual Review of Public Health*, *17*(1), 47–67. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8724215http://www.annualreviews.org/doi/10.1146/annurev.pu.17.050196.000403 doi: 10.1146/annurev.pu.17.050196.000403

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*(6), 1481–1496. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03610929708831995 doi: 10.1080/03610929708831995

Kulldorff, M. (2015). *SaTScan User Guide V9.4*.

Lanfranchi, A. (2015). Induced abortion and breast cancer. *Law & Medicine*, *30*(2), 143.

Luginaah, I. N., Gorey, K. M., Oiamo, T. H., Tang, K. X., Holowaty, E. J., Hamm, C., & Wright, F. C. (2012). A geographical analysis of breast cancer clustering in southern Ontario: Generating hypotheses on environmental influences. *International Journal of Environmental Health Research*, *22*(3), 232–248. doi: 10.1080/09603123.2011.634386

Lynch, H. T., Marcus, J. N., Watson, P., & Lynch, J. F. (1989). Familial and Genetic Factors New Evidence. In B. A. Stoll (Ed.), *Women at high risk to breast cancer* (pp. 27–39). Springer Netherlands. doi: 10.1007/978-94-009-1327-1_3

McPherson, K., Steel, C. M., & Dixon, J. M. (2000). ABC of Breast Diseases Breast cancer epidemiology , risk factors , and genetics Risk factors for breast cancer. *Br. Med J.*, *321*(September), 624–628.

Middlesex Economic Development. (2018). *Our Community*. Retrieved 2018-06-30, from https://www.investinmiddlesex.ca/our-community

Mills, P. K., & Yang, R. (2006). *Regression Analysis of Pesticide Use and Breast Cancer Incidence in California Latinas* (Vol. 68) (No. 6). Retrieved 2017-11-02, from http://0-search.ebscohost.com.library.unl.edu/login.aspx?direct=true{\&}db=a2h{\&}AN=19570059{\&}site=ehost-live{\&}scope=site

Ozdenerol, E., Williams, B. L., Kang, S. Y., & Magsumbol, M. S. (2005, aug). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *International journal of health geographics*, *4*(1), 19. Retrieved from http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-4-19 doi: 10.1186/1476-072X-4-19

Pan, S. Y., Morrison, H., Gibbons, L., Zhou, J., Wen, S. W., DesMeules, M., & Mao, Y. (2011, may). Breast Cancer Risk Associated With Residential Proximity to Industrial Plants in Canada. *Journal of Occupational and Environmental Medicine*, *53*(5), 522–529. Retrieved from http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage{\&}an=00043764-201105000-00010 doi: 10.1097/JOM.0b013e318216d0b3

Pong, R. W., DesMeules, M., & Lagacé, C. (2009, feb). Rural-urban disparities in health: How does Canada fare and how does Canada compare with Australia? *Australian Journal of Rural Health*, *17*(1), 58–64. Retrieved from http://doi.wiley.com/10.1111/j.1440-1584.2008.01039.x doi: 10.1111/j.1440-1584.2008.01039.x

Reynolds, P., Hurley, S. E., Goldberg, D. E., Yerabati, S., Gunier, R. B., Hertz, A., ... Ziogas, A. (2004, oct). Residential proximity to agricultural pesticide use and incidence of breast cancer in the California Teachers Study cohort. *Environmental Research*, *96*(2), 206–218. Retrieved from https://www.sciencedirect.com/science/article/pii/S0013935104000520 doi: 10.1016/J.ENVRES.2004.03.001

Robert, S. A., Strombom, I., Trentham-Dietz, A., Hampton, J. M., Mcelroy, J. A., Newcomb, P. A., & Remington, P. L. (2004). Socioeconomic Risk Factors for Breast Cancer: Distinguishing Individual-and Community. *Source: Epidemiology*, *15*(4), 442–450. Retrieved from https://www-jstor-org.proxy1.lib.uwo.ca/stable/pdf/20485927.pdf?refreqid=search{\%}3A391fc0ee02faa7d900aa9b2626d0877f doi: 10.1097/01.ede.0000129512.6169

Schettler, T. (2013). The Ecology of Breast Cancer. *Science & Environmental Health Network*(October), 1–200.

Statistics Canada. (2011). *2011 Census Program Data Products.*

Statistics Canada. (2017). *Census dictionary.* Retrieved 2017-11-05, from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

Statistics Canada. (2018). *The 2011 National Household Surveythe complete statistical story.* Retrieved 2018-10-20, from https://www.statcan.gc.ca/eng/blog-blogue/cs-sc/2011NHSstory

Tobler, W. R. (1970, jun). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234. Retrieved from http://www.jstor.org/stable/143141?origin=crossref doi: 10.2307/143141

Trichopoulos, D., Hsieh, C.-C., Macmahon, B., Lln, T.-M., Lowe, C. R., Mirra, A. P., … Yuasa, S. (1983). Age at any birth and breast cancer risk. *International Journal of Cancer*, *31*(6), 701–704. Retrieved from http://doi.wiley.com/10.1002/ijc.2910310604 doi: 10.1002/ijc.2910310604

Weiderpass, E., Meo, M., & Vainio, H. (2011, mar). Risk factors for breast cancer, including occupational exposures. *Safety and health at work*, *2*(1), 1–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22953181http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3431884 doi: 10.5491/SHAW.2011.2.1.1

Ye, Z., Gao, D. L., Qin, Q., Ray, R. M., & Thomas, D. B. (2002). Breast cancer in relation to induced abortions in a cohort of Chinese women. *Br J Cancer*, *87*(9), 977–981. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12434288{\%}5Cnhttp://www.nature.com/bjc/journal/v87/n9/pdf/6600603a.pdf doi: 10.1038/sj.bjc.6600603

4 Conclusions

This last chapter summarizes the findings of the study and describes the contribution that this thesis can provide. The strengths and limitations of the study are also discussed. Lastly, we provide recommendations for future work in breast cancer research.

4.1 Summary of Findings

We utilized GIS tools and spatial analysis to explore breast cancer prevalence and detect clusters in Middlesex County and its surrounding areas. The analysis reveals spatial patterns reflected by the data and the results show the breast cancer cluster locations, which are located on both western and eastern fringes of the county. Using regression analyses, the study provides evidence of significant socioeconomic variables that elevate the risk, including average income, employment rate, and women's occupation in agriculture, forestry, fishing and hunting.

4.2 Research Contributions

To the best of our knowledge, a study to locate breast cancer clusters like ours has not been reported for Middlesex County. The results show persistent breast cancer cluster locations for both high and low values, and also the combined influence of socioeconomic and environmental risks for breast cancer in the study area, such that certain geographic areas may need specific policy attention.

One of the challenges for public health agencies when dealing with cancer clusters is to communicate with the public effectively when concerns about the clusters arise. With the involvement of the MLHU in the study, the results may contribute to a knowledge exchange with the community to not only better understand breast cancer and its link to environmental health, but also contribute to the programs provided by the health unit to advance breast cancer prevention.

4.3 Strengths and Limitations

We recognize some limitations that constrain our study and the interpretation of results. As mentioned in Section 3.4, the use of census data is subject to miscount or under-enumeration because they are estimates and this study used them to detect breast cancer clusters both at the point and aggregated data. The census data includes population count and socioeconomic factors. An ideal analysis would be conducted with precise and comprehensive data to produce high-quality results, but in reality, the availability of such data are rare. Breast cancer prevalence for the eleven years of observation was measured using intercensal data in 2006, 2011, and 2016 and even though this method is widely used in various cancer studies, there is inaccuracy in the population registry that may affect the results.

Another limitation of studies that use census data is the frequent changes of administrative borders that can cause discrepancies including area size, population count for that area and the values of socioeconomic variables yielded from the area. It is not a problem for large-scale studies, but careful interpretation must be implemented to their small counterparts.

Furthermore, health data is subject to errors such as diagnostic error or misclassification. There could be double counting or under registration in the cancer registry. This kind of error might not happen often, but it might encourage false findings if it happened in an area with cluster tendencies *a priori*.

Regardless of the limitations, this study offers many strengths in data quality and the methods it uses. The availability of breast cancer data at the individual level is powerful because, with such granularity, we were able to conduct point data analysis using SaTScan and then there was room for the study to be expanded at different levels of geographical unit aggregation. Not only were breast cancer cases analyzed as points and as prevalence per area, the study area was also expanded to its land-locking counties to anticipate edge effects in pattern analysis of the single county. All results were cross-checked and the use of different methods provides more confidence in the findings.

4.4 Direction for Future Research

Based on the results that report locations of breast cancer in the county where prevalences are either high or low and significant socioeconomic factors increase the risk, future research may be conducted to explore the subject further. It is important to examine contextual determinants of the cluster in-depth to better understand risk factors of breast cancer.

Further analysis that incorporates the distance to the screening programs and cancer care centres may help to explore the association between residential location and an increased risk of breast cancer. Our result shows the clusters are located mostly in the rural areas where there is less accessibility to the cancer care facilities. The inclusion of this factor may better explain the geographical pattern of breast cancer occurrences.

Given the challenges of cancer studies related to the latency of breast cancer development, a qualitative analysis involving interviews with breast cancer patients or survivors within the identified clusters may help to gain more insights about breast cancer. The information collected may include family history, work and residential history, exposure to chemicals at home or workplace, and other contextual environmental exposure variables. On a larger scale, a quantitative analysis in the form of a case-control community health survey might also provide a better understanding of the correlation between the identified clusters and breast cancer.

Not only does personal history play a role in breast cancer development, but location history may also provide etiological clues to better understand breast cancer risk factors. A historical analysis within the identified clusters may be useful to draw a time-lapse conceptual and environmental exposure ideas to locate carcinogens. The analysis may utilize historical maps, city directories, industrial and business directories that can potentially capture contaminated areas within the clusters that could lead to more breast cancer etiological clues.

Appendix A Regression parameters

The 31 variables used in the regression analysis were retrieved from The Canadian Census Analyser that provided access to Canadian Census Data (http://datacentre.chass.utoronto.ca/census/). The description of all variables was directly quoted from this source. Models for the analysis are listed below in R syntax.

Variable	Description
AVG_INC	Income of individuals in 2010 (part 1) - Females / Total income in 2010 of population aged 15 years and over; Females / Average income; Females
NOCERT	A total value of two variables: - Education - Females / Total population aged 25 to 64 years by highest certificate diploma or degree; Females / No certificate diploma or degree; Females - Education - Females / Total population aged 25 to 64 years by highest certificate diploma or degree; Females / High school diploma or equivalent; Females
CERT	Education - Females / Total population aged 25 to 64 years by highest certificate diploma or degree; Females / Postsecondary certificate diploma or degree; Females
EMP_RATE	Labour force status - Females / Employment rate; Females
ETH_ABO	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / North American Aboriginal origins; Females
ETH_OTH	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / Other North American origins; Females
ETH_EUR	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / European origins; Females
ETH_CAR	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / Caribbean origins; Females
ETH_LAT	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / Latin Central and South American origins; Females
ETH_AFR	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / African origins; Females
ETH_ASIA	Ethnic origin population - Females / Total population in private households by ethnic origin; Females / Asian origins; Females
IND11	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 11 Agriculture forestry fishing and hunting; Females
IND21	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 21 Mining quarrying and oil and gas extraction; Females
IND22	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 22 Utilities; Females

IND23	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 23 Construction; Females
IND31	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 31-33 Manufacturing; Females
IND41	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 41 Wholesale trade; Females
IND44	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 44-45 Retail trade; Females
IND48	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 48-49 Transportation and warehousing; Females
IND51	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 51 Information and cultural industries; Females
IND52	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 52 Finance and insurance; Females
IND53	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 53 Real estate and rental and leasing; Females
IND54	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 54 Professional scientific and technical services; Females
IND55	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 55 Management of companies and enterprises; Females
IND56	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 56 Administrative and support waste management and remediation services; Females
IND61	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 61 Educational services; Females
IND62	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 62 Health care and social assistance; Females
IND71	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 71 Arts entertainment and recreation; Females

IND72	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 72 Accommodation and food services; Females
IND81	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 81 Other services (except public administration); Females
IND91	Industry - Females / Total labour force aged 15 years and over by industry - North American Industry Classification System (NAICS) 2007; Females / All industries; Females / 91 Public administration; Females

Full model:

glm($meanaar \sim AVG\_INC + NOCERT + CERT + EMP\_RATE + ETH\_ABO + ETH\_OTH + ETH\_EUR + ETH\_CAR + ETH\_LAT + ETH\_AFR + ETH\_ASIA + IND11 + IND21 + IND22 + IND23 + IND31 + IND41 + IND44 + IND48 + IND51 + IND52 + IND53 + IND54 + IND55 + IND56 + IND61 + IND62 + IND71 + IND72 + IND81 + IND91$, weights = $population$)

Model with 98% PCA for 2006 data at the CSD level (5 principal components):

glm($meanaar \sim EMP\_RATE + AVG\_INC + IND11 + IND22 + IND21$, weights = $population2006$)

Model with 99% PCA for 2011 data at the CSD level (5 principal components):

glm($meanaar \sim EMP\_RATE + AVG\_INC + IND11 + IND22 + IND21$, weights = $population2011$)

Model with 70% PCA for 2006 data at the DA level (10 principal components):

glm($meanaar \sim IND11 + AVG\_INC + EMP\_RATE + IND51 + IND71 + ETH\_ABO + IND48 + IND23 + IND41 + IND53$, weights = $population2006$)

Model with 70% PCA for 2011 data at the DA level (11 principal components):

glm($meanaar \sim IND11 + EMP\_RATE + IND71 + AVG\_INC + IND23 + ETH\_ABO + IND51 + IND41 + IND48 + IND53 + NOCERT$, weights = $population2011$)

Appendix B SatScan parameters and values

The analysis was run with SatScan software using the following parameter values:

Section	Parameter	Value
Input	Case File	a comma-separated values (CSV) file with a list of areas and case counts
	Population File	a CSV file with a list of areas and population counts
	Coordinates File	a CSV file with a list of areas and X and Y coordinates
	Time Precision	Year
	Study Period: Start Date	2003
	Study Period: End Date	2013
	Coordinates	Lat/Long
Analysis	Type of Analysis	Space-Time
	Probability Model	Discrete Poisson
	Scan For Areas With	High or Low Rates
	Time Aggregation	Year
	Length	1 Year
Advanced Analysis	Maximum Spatial Cluster Size	varies from 1% up to 30%
	Spatial Window Shape	Circular
	Inference: Monte Carlo Replications	999
Output	Text Output Format	an output text file
	Geographical Output Format	(checked) Shapefile for GIS Software
	Column Output Format	(all five checkboxes were checked for dBase)

Curriculum Vitae

Name:	Jenny Tjhin
Post-Secondary Education and Degrees:	Parahyangan Catholic University Bandung, Indonesia 1998-2004 B.Sc. University of Western Ontario London, ON 2016-2018 M.Sc.
Honours and Awards:	SSHRC CGS M 2017-2018
Related Work Experience:	Research Assistant University of Western Ontario 2016 - 2017 Teaching Assistant University of Western Ontario 2016 - 2018