Final Report: Assessment of the presence and magnitude of Modifiable Areal Unit Problem at various levels of geographic data aggregation in Ontario Health Central

By

Anastasiia Smirnova (500940333)

Ellie Maclennan (501081829)

Inderjeet Punia (500885851)

Nating Zhao (501210300)

Sigao Li (500844153)

Submitted

To

OCHPP Team at MAP-CUHS

St. Michael's Hospital

&

Professor Lu Wang

Geography & Environmental Studies

In partial fulfillment of the requirements

for

SA 8904 - GIS Project Management Applications

December 1, 2022

Toronto Metropolitan University

Table of Contents

Executive Summary

This report was prepared for the client, the Ontario Community Health Profiles Partnership (OCHPP) team at the MAP Centre for Urban Health Solutions at St. Michael's Hospital. The aim of this project was to assess the presence and magnitude of the Modifiable Areal Unit Problem (MAUP), the problem of different spatial units yielding different results, at various levels of geographic data aggregation in Ontario Health Central. Recently, new "neighbourhood" boundary regions, the smallest spatial units available, were created in Toronto, and this will be expanded for other Ontario regions. In addition to neighbourhoods, there are multiple administrative regional boundaries in Ontario used in health research, such as various types of Census regions or provincial boundaries. It follows that health research in Ontario would also be impacted by the MAUP, but the extent of these effects tend to be overlooked and are largely unstudied in Ontario.

This study used disease data provided by the OCHPP and census data from Statistics Canada to conduct Multivariate Regression Analysis and Predictive Analysis on 366 neighbourhoods in southern Ontario. The results indicate that significant predictors at different geographic levels are distinct, and that a model from a high aggregation level will not be able to accurately predict the dependent variable at a low aggregation level, but vice versa. Furthermore, as the levels of aggregation increase, there is only an increase in correlation between a few variables and others being weakened.

Thus, this study provides support that the MAUP influences diabetes research when using different levels of geographic data aggregation. In this project's findings, some variables do not remain significant at the CSD level. This was contradictory to typical findings in the literature seeing stronger correlations with higher data aggregation, but this finding is likely attributed to a small sample size and consequent lack of power in the CSD analyses. However, and fortunately for the sake of confidence in the diabetes literature, the overall correlations between the predictors and diabetes prevalence remained generally consistent. We further discuss limitations and recommendations to handle the MAUP.

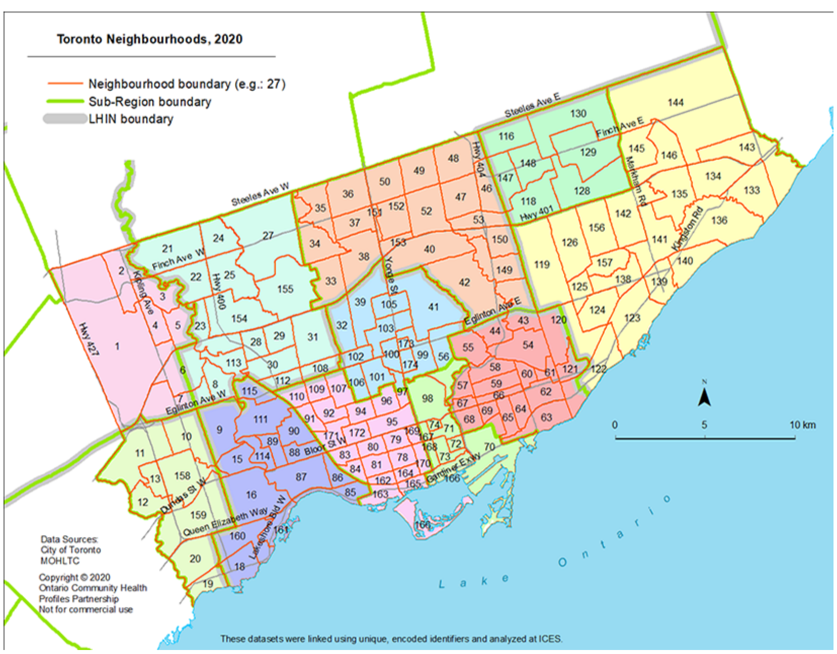
Introduction

Research has shown a significant influence of place of residence on health. Several studies that focused on the relationship of neighbourhoods and health in particular, confirmed that physical and socioeconomic factors associated with neighbourhoods, including social program availability, services, infrastructure and facilities, access to green space and other indicators, can impact health in both positive and negative ways (Awuor & Melles, 2019; O′Campo et al., 2015; Vallée et al., 2020). Health may also be directly affected by good access to a general practitioner or other medical service provider, even if the medical facilities are not located in the neighbourhood itself, good transport links may still provide residents with an advantage in terms of their health.

In addition, administrative boundaries are often used in health research, often because these are the regions with available data. Common boundaries used in health research come from Canadian Census geographic areas (e.g., Census subdivisions), as well as Local Health Integration Networks (LIHNs), Sub-Regions, and neighbourhoods (see Figure 1).

Figure 1

Common Administrative Health Boundaries in Ontario



Note: Taken from the Ontario Community Health Profiles Partnership website.

Much of the health data available in Ontario, similar to other jurisdictions, is provided at the neighbourhood level. In Ontario, neighbourhoods are the smallest available spatial unit, and thus, are often the scale often used in health-related research. Recently, the previous Toronto neighbourhood boundaries (n=140) were replaced with new neighbourhood boundaries (n=158)(See Appendix B).

One of the problems of relying heavily on available administrative data in research is associated with the MAUP, an issue related to the effects of scaling and zoning of spatial units in the field of spatial analysis (Openshaw and Taylor, 1979). Doing analysis on selected administrative boundaries, often used out of necessity, impacts the results and interpretations, though these effects often go without significant consideration. Thus, researching the impact of MAUP in the context of Toronto and its surrounding areas is necessary. Due to the change to the neighbourhood regions, and their being the smallest unit available and consequent frequent use in health research, neighbourhoods are the primary geographic area of focus to assess for this project.

Some of the commonly used indicators for measuring neighbourhood related population health outcomes include premature mortality, life expectancy and chronic conditions, including diabetes (Awuor & Melles, 2019; Gariepy et al., 2015). Meanwhile, diabetes, a chronic disease, is one of the leading causes of death in Canada (Government of Canada, 2022). Consequently, this paper aims to use diabetes as a case study, contribute to the knowledge base on the prevalence of diabetes, and consider the influence of MAUP and its impacts on the distribution of diabetes cases and its predicting factors in selected regions of Southern Ontario at the neighbourhood versus census subdivision (CSD) level. This study explores how great these differences are, and what impact they may have in findings, interpretation, and policy. Further, this project aims to assess if results at different data aggregations correspond to existing findings about predictors of diabetes in the literature and evaluate if novel variables influence the prevalence of diabetes in Toronto and the neighbouring areas. Finally, based on the findings, recommendations are made for researchers in working with health data in Ontario.

Hypotheses

1. Significant predictors will be distinct at different geographical levels.

2. Stronger correlations between diabetes and independent variables are expected as the level of aggregation increases.
3. A model from one level of aggregation will not accurately predict diabetes prevalence at a different level of aggregation.

Literature Review

The Modifiable Areal Unit Problem

The Modifiable Areal Unit Problem (MAUP), coined by Openshaw and Taylor (1979), was first discussed through investigations of how data aggregation impacted correlation coefficient values. Here, the authors found their correlation coefficient values changed with changes in boundaries and spatial data aggregation. Notably, greater levels of data aggregation tended to see stronger correlations. This highlighted two fundamental components of what is now known as the MAUP. First, the often arbitrary and typically administrative drawing of boundaries can yield differing results (zoning). Second, the scale at which data are analyzed may also see varied outcomes (scaling).

Subsequent research in geostatistics echoed these findings (Openshaw, 1984; Cressie, 1996). Fotheringham and Wong (1991) discussed the MAUP in multivariate statistical analysis (here, being regression models), and voiced concern regarding their findings of scale and zoning yielding unpredictable impacts on results. They found both differences in intensity and effects of analyses, such that even negative correlations could become positive at different levels of scale.

Since then, the impacts of the MAUP has been documented across a range of disciplines, such as conservation (e.g., Moat et al., 2018), public safety (e.g., Xu et. al 2018), politics (e.g., Lee & Rogers, 2019), and business (e.g., Cartone & Postiglione, 2019). Similarly, the MAUP also plays an important role in the field of public health. For example, a study conducted using data from the region of Picardy, France, explored this phenomenon. This study employed three administrative spatial scales (the smallest available units, grid cells composed of squares of the same sizes, and counties of irregular sizes and shapes) to explore the relationship between exposure indicators, socioeconomic factors, and health outcomes (Saib et al., 2014). Through this research, the authors found consistent results across the three scales on some measures, though they noted stronger associations in their larger geographical units. Similar to earlier studies (e.g., Openshaw and Taylor, 1979), these authors attributed their finding of stronger associations with greater areas to data aggregation. Fortunately, in this study, the authors still found generally consistent results across their aggregations, minimizing the magnitude of the MAUP in this case. Further, their exposure variables saw the least amount of variability across the spatial scales, while their composite socioeconomic variable saw the greatest variability, highlighting the ranging and uncertain impacts the MAUP poses.

More recently, in response to numerous studies pointing to environmental associations with COVID-19 outbreak data, one study examined the impact of MAUP on such associations between environmental factors (here using NO_2) and COVID-19 mortality using data from two Chinese provinces, Henan and Hubei, as a case study (Wang & Qian, 2020). Here, the authors found the associations between COVID-19 deaths and NO_2 varied across both aggregation level and strategy, leading the authors to encourage caution in conducting and using geographic findings related to COVID-19 in order to better guide public health measures.

MAUP is often disregarded or referred to as unsolvable in research. However, some scholars have attempted to solve or decrease the impact of MAUP in different ways. For instance, in addition to a comparison of statistical models and their results at different scales, a number of studies focusing on income segregation (i.e., separation of various classes of people by income) tried to tackle MAUP by conducting a multi-level analysis which uses a model-based approach that investigates spatial effects at multiple scales simultaneously (Jones et al., 2018; Quick and Revington, 2022; Johnston, 2016). This approach analyzes variance at one scale while excluding the variation at other scales. As a result, according to Jones et al. (2018), while most preceding research showed the biggest segregation at the finest scale with its measured intensity declining at higher spatial scales, several studies using a

Multilevel approach have discovered that there was a bigger segregation at a higher scale (Jones et al., 2018; Quick and Revington, 2022).

In the context of Toronto, a multilevel analysis study focusing on income segregation, that analysed MAUP using data at three levels of aggregation (Census Tract (CT), dissemination area (DA), and neighbourhood), defined a few interesting patterns (Quick & Revington, 2022). The scholars discovered more income segregation within the Neighbourhood Improvement Areas (NIAs), while in the city centre there was less segregation. On a bigger scale, the city centre showed very different results from the rest of the city.

Another Toronto-specific example of tackling MAUP by Hazell and Rinner (2020) examined the use of area based composite indices to model and evaluate urban environmental conditions of butterfly populations across the city at two scales - dissemination areas and census tracts. The researchers concluded that “the scale at which the data are aggregated had a greater impact on the overall model fit compared to the composite indexing approach, whereby the CT-level models generally performed better than the DA-level models” (Hazel and Rinner, 2020, p.1674). All Toronto based studies under analysis admitted the influence of scale on the results and the presence of zoning and scaling effects on the results (Hazell & Rinner, 2020; Mitra & Buliung, 2012; Quick & Revington, 2022).

Diabetes Prevalence and Predictors

There is considerable evidence to show that socio-economic status (SES) and its constituent elements are associated with determinants of health. Diabetes as one of cardiovascular diseases, shows a significant socio-economic gradient in the prevalence of disease risk factors (Rabi et al ,2006). Education, income, race, and immigration status are found to be significantly associated with diabetes prevalence. Diabetes prevalence was higher among individuals with lower income, fewer educational qualifications, and non-professional occupations, those with lower SES are more likely to develop diabetes and suffer from worse outcomes.

Years of education and income as important principles of measuring SES were also selected as indicators to computed deprivation index in the research of Tompkins, et al (2010). In this study, an analysis of the patterns of correspondence between high diabetes rates and socioeconomic determinants of health was conducted by overlaying diabetes prevalence rates and principal components, the result shows that the principal components explained the higher percentage of variance were referred to as low income, high rental, unemployed, low education, lone parent, and visible minority. Agardh et al (2011) used the methods applied in the comparative risk assessment to explore the association between lower educational levels and type 2 diabetes incidence and concluded that there is a considerable burden of type 2 diabetes attributed to lower educational levels in Sweden. A number of studies have also demonstrated that low-income populations are more likely to develop diabetes. Rabi et al (2006) generated household income quintiles from DA annual income data and found that the lowest quintiles have the highest rates of referral and also higher rates of diabetes than the upper quintiles.

In terms of diabetes prevalence and incidence, racial and ethnic disparities are an important public health issue. As a consequence of the racial and socioeconomic patterns of segregation, obesity and type II diabetes are theorized to be influenced by disparities in neighbourhood environments. As compared to White participants, Black and Hispanic participants had 2.89 times and 1.48 times the odds of developing T2DM (Piccolo et al. 2015). In bivariate analyses, there was a positive association between the prevalence of diabetes and the percentage of non-Hispanic black and Hispanic residents in Washington (Piccolo et al., 2015). Both Toronto and Chicago experience an association between ethnic groups and diabetes rates that is closely related to immigration trends, (Kolpak and Wang, 2017).

Diabetes prevention must consider factors about individual-level behavioral lifestyle like physical activity. However, upstream environmental factors like the urban built environment is a growing recognition as potential targets for intervention. Many studies have focused on the neighbourhood's context which can affect the health of individuals. Walking-friendly neighbourhoods, easy access to services, and a variety of transit options can promote physical activities such as walking and bicycling (Awuor & Melles, 2019). There is a strong and consistent association between the availability of walkable destinations and transportation behaviors and diabetes. The distribution of parks and other green spaces can influence the frequency and intensity of physical activity. Piccolo et al. (2015) indicate that access to parks and green space may potentially reduce diabetes.

There have already been several studies examining differences in diabetes prevalence according to individual SES. However, any study that examines the relationship between health and place will be influenced by the scale and design of the zoning used in the study. The MAUP can have a significant effect on the analytical results of the same input data collected under different spatial units. A few health studies have focused on MAUP, this study will investigate diabetes prevalence and SES indicators in the Ontario region using two different spatial units in order to explain how the MAUP can affect results.

Data and Methodology

Study Area

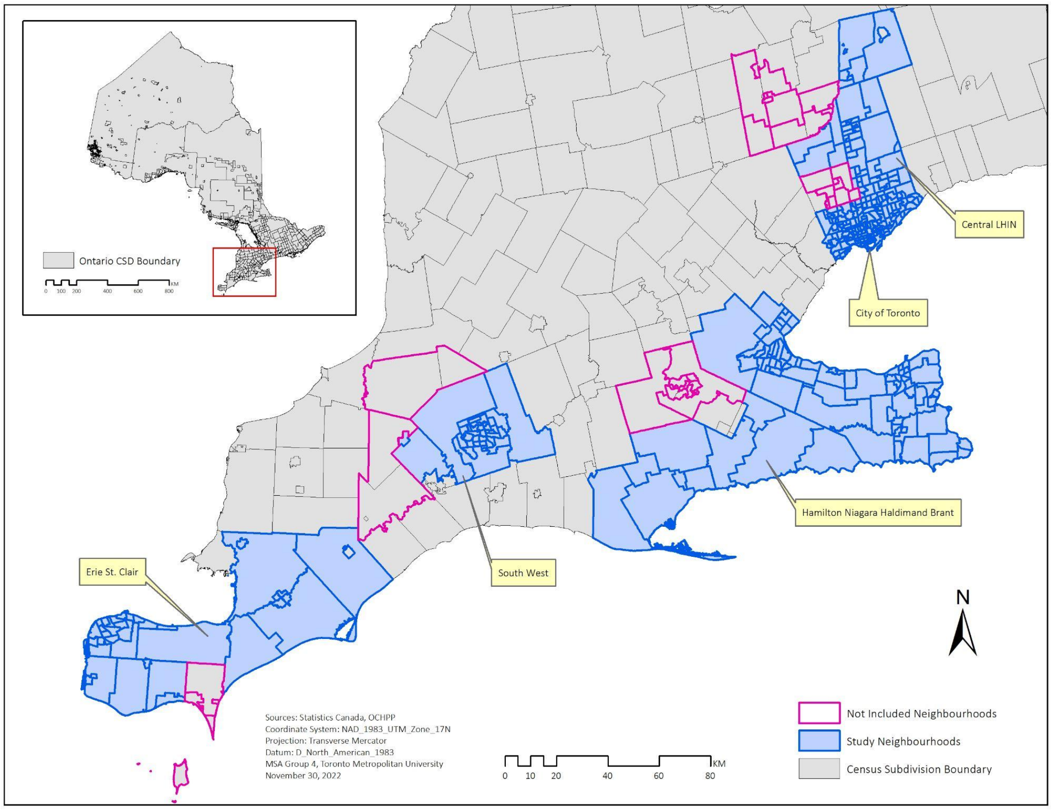
In Canada, there are already more than 2.3 million people with diabetes aged 18 and over in 2021, nearly half of whom reside in Ontario (Statistics Canada, 2022). Southern Ontario, as the main gathering area of Ontario's population, is an excellent research area, and many health-related organizations regularly publish neighbourhood-level health data in this area.

This study first used neighbourhood-level data to analyze the demographic distribution of diabetes and used the Census Subdivision (CSD) for MAUP comparisons. The study area is located in southern Ontario, which are distributed among five areas: City of Toronto, Central LHIN, Hamilton Niagara Haldimand Brant, South West, and Erie St. Clair.

Disease data used in this study were obtained from the OCHPP. In the OCHPP system, there are 396 neighbourhoods, but due to the particularity of the MAUP study, 30 neighbourhoods were excluded because the boundaries were not aligned with the CSD, across multiple CSDs, or were assigned to different CSDs (see Figure 2). In the end, 366 neighbourhoods were included in the study, and these neighbourhoods were formed into 37 Census Subdivisions after aggregation. Within the study area, the total population of 20+ is about 5 million, and people with diabetes over the age of 20 accounts for about 12% (0.6 million) of the total population.

Figure 2

Map of Study Area



Variables

The data were divided into two sections: disease data from OCHPP and 2016 census data from Statistics Canada (see Table 1). This study's primary disease outcome measure was the number of diabetes cases reported by the OCHPP among residents aged 20+, The included variables are selected on the basis of following criteria: acknowledged association to health, not too specific or too broad, have high likelihood of causality, availability in the 2016 Canadian census data, and the most important is their established relationships as socioeconomic determinants of health in the literature. In this study, we choose median after-tax income of households, renter and unemployment as indicators of income, and no certificate, diploma, or degree to assess the education level, and immigration status and visible minority variables are also included to measure the SES. As for physical activity, we choose to use the main mode of commute by bike or walk as a variable.

Table 1

Study Variables

Name	Variables Name	Year	Spatial Resolution	Data Source	Description
Adult Health and Disease-Diabetes	Per_DC_Both	2019	Neighbourhood	Ontario Community Health Profiles Partnership	Number of people with diabetes 2018/19, All Ages 20+
Ontario Conversion File	/	2016	Disseminations Areas, Neigbourhood		Provided by OCHPP to aggregate data from the Disseminations Areas level to other higher levels
Median after-tax income of households in thousand ($)	MATI_HH_K		Disseminations Areas	Statistics Canada 2016 Census	Number of after-tax income recipients aged 15 years and over in private households
Education - No certificate, diploma, or degree (%)	Per_NO_CDD				Highest certificate, diploma or degree for the population aged 15 years and over in private households
Visible minority in private households (%)	Per_VM				Visible minority for the population in private households
Immigration in private households (%)	Per_Imm				Immigrant status and period of immigration for the population in private households
Unemployment (%)	Per_Unemp				Population aged 15 years and over by Labour force status
Renter (%)	Per_Renter				Private households by tenure
Physical activity - Main mode of commute by bike or walk (%)	Per_JtW_WB				Main mode of commuting for the employed labour force aged 15 years and over in private households with a usual place of work or no fixed workplace address

Data

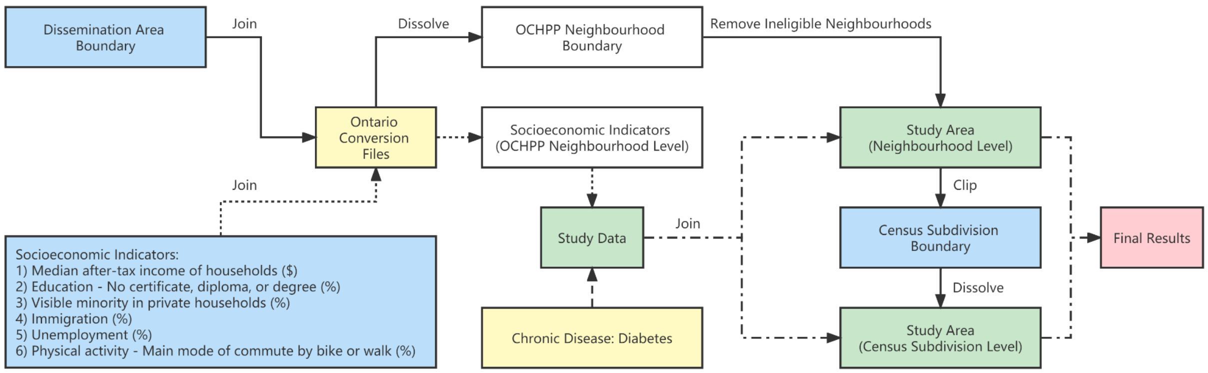
The initial step is to create the neighbourhood boundary, GIS software Arcmap was utilized to generate boundary files for the research. Since OCHPP and Ontario do not share the same definition of neighbourhood, GIS software is required to create the boundary. Download the DA and CSD boundary data for Ontario from Statistics Canada. Then, through the Ontario Conversion File, execute "Join," provide DA with the neighbourhood ID where it is located, and then execute "Dissolve" to obtain the OCHPP neighbourhood boundary. After removing neighbourhoods that are ineligible for MAUP studies, "Clip" yields the same study area at the CSD level.

The second step is to collect socioeconomic and disease data. The 2016 census data were obtained from Statistics Canada, and the data at the DA level that synthesized the variables above were chosen. Since this study needs to be analyzed at two levels, neighbourhood and CSD, all DA data must be aggregated, which requires using OCHPP's Ontario Conversion Files. All DA neighbourhood IDs can be assigned to DAs by merging, followed by a merging sum. Due to the fact that the Median after-tax income of households is not a count value, it must be multiplied by the number of households in each DA and then divided by the number of households in the area after being summed at the neighbourhood or CSD level. Finally, perform percentage processing on variables other than the Median after-tax household income. The OCHPP website provides chronic disease data for 2018-2019 for all neighbourhoods, all data can be downloaded directly. The diabetes data are extracted from the downloaded file, but since the file is in.xlsx format, a separate.xls file will be created for diabetes in order to manage these data efficiently, and the tables will be merged according to the OCHPP Neighbourhood ID.

The final step involves combining processed socioeconomic data and disease data with boundary files (see Figure 3), followed by multivariate regression analysis in SPSS, predictive analysis using the regression results, and visualization in Arcmap.

Figure 3

Data Processing Workflow



Method

Step One:

Data created from the figure above will be loaded into SPSS and rated. Once rated, it will allow the diabetes variable to be mapped for NH and CSD levels.

Step Two:

The models will be created on SPSS using the linear regression analysis tool. The stepwise function will be used in order to ensure the creation of a statistically sound model that passes all multicollinearity tests and manual revision will be done. Stepwise regression removes all variables that are insignificant or weakly correlated, thus going through many models in order to find the one with the highest R^2 value.

Step Three:

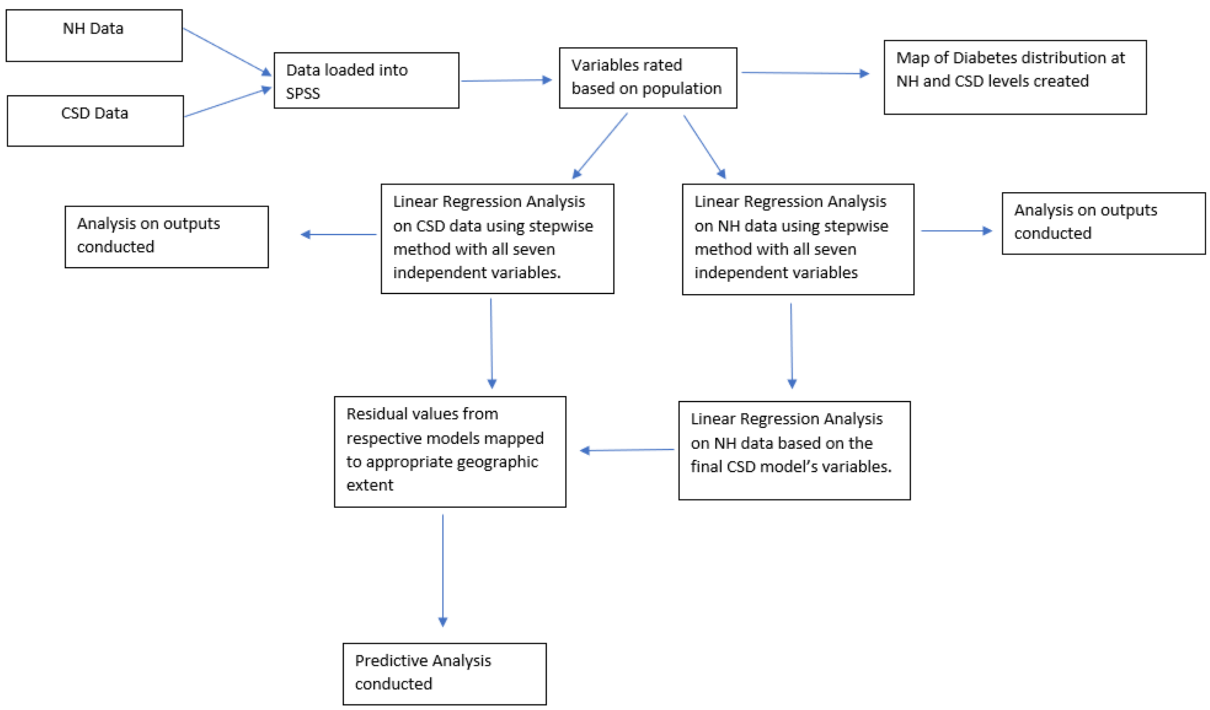
Another model will be created on the NH level to reflect the variables chosen by the stepwise process in the CSD model. These two models will then have different geographical scales but the same variables. The residual values for both the models will be mapped and compared afterwards in order to highlight the MAUP effect.

Step Four:

The predictive analysis will be performed, using the regression coefficients at the neighbourhood and CSD levels as predicted values, and "predicting" at another level. The predicted results are then compared with the actual results to see how the predictions differ between the two levels.

Figure 4

Method Workflow

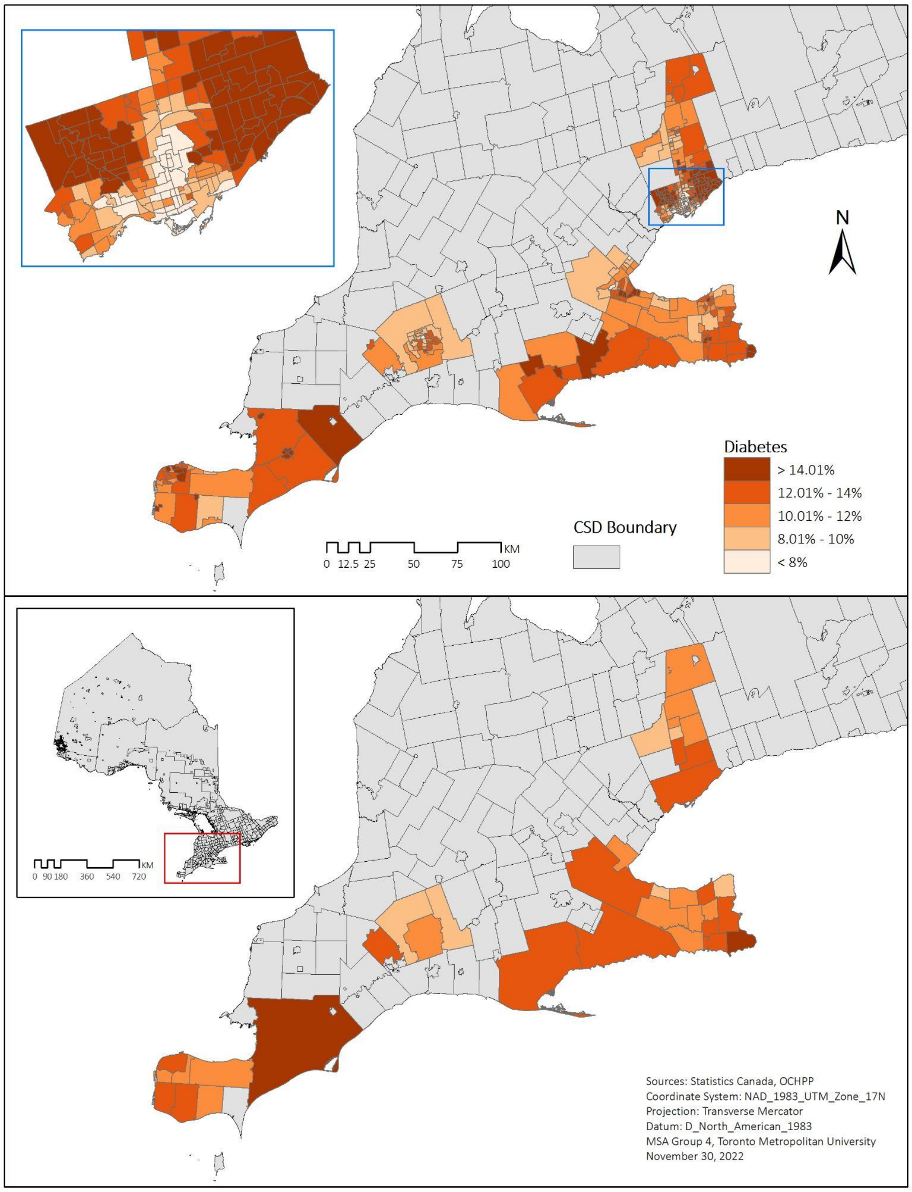


Data Analysis and Interpretation

In order to portray MAUP intuitively, this study employs the same diabetes range mapping at the neighbourhood and CSD levels (see Figure 5). Northwest and East Toronto had a high prevalence of diabetes at the neighbourhood level, but significantly lower rates at the CSD level. This is due to the merger of Toronto's at the CSD level. However, there are few diabetics in central, southern, and southwestern Toronto; hence, when the statistics are averaged at the CSD level, it appears that Toronto has fewer diabetes overall. Similar phenomena occurred in several other neighbourhoods, including those to the east of Erie St. Clair, the south of Hamilton Niagara Haldimand Brant, and some other locations.

Figure 5

Map of Diabetes Distribution at the Neighbourhood and CSD levels



Multivariate Regression Analysis

Multivariate regression helps in determining relationships and analysing patterns with large data sets. Due to the nature of MAUP, one of the ways to compare variables between geographical regions is to create models and see differences in variable relationships.

The first model created will be based on the 37 census subdivisions and the second model will be based on 366 neighbourhoods. The dependent variable that will be used across both models is total diabetes in 2018/2019 for ages 20+. By creating two different models with different data aggregations on the same geographical area, the extent of the MAUP is displayed.

This study is able to see from the models how certain variables are significant in the NH model versus the CSD model. The analysis starts by adding variables to the model; if they are not significant, SPSS removes them. Table 2 and Table 3 show the models created. From the models this study can see that 6 out of the 7 independent variables are significant in the neighbourhood model, but only 3 out of 7 are significant in the CSD model.

Table 2

CSD Regression models

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square	F Change	Sig. F Change	Durbin-Watson
1	.691[a]	0.478	0.463	1.114714035	0.478	32.066	0.000	
2	.756[b]	0.571	0.546	1.025203353	0.093	7.379	0.010	
3	.799[c]	0.639	0.606	0.954581986	0.068	6.217	0.018	2.295

a. Predictors: (Constant), MATI_HH_K
b. Predictors: (Constant), MATI_HH_K, Per_VM
c. Predictors: (Constant), MATI_HH_K, Per_VM, Per_Imm
d. Dependent Variable: Per_DC_Both

Table 3

NH Regression models

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square	F Change	Sig. F Change	Durbin-Watson
1	.700[a]	0.491	0.489	2.348323766	0.491	350.673	0.000	
2	.792[b]	0.628	0.625	2.010914867	0.137	133.398	0.000	
3	.838[c]	0.702	0.700	1.800252602	0.075	90.926	0.000	
4	.855[d]	0.732	0.729	1.711619121	0.029	39.462	0.000	
5	.863[e]	0.745	0.741	1.671899592	0.013	18.356	0.000	
6	.871[f]	0.759	0.755	1.627472963	0.014	20.923	0.000	1.127

a. Predictors: (Constant), Per_NO_CDD
b. Predictors: (Constant), Per_NO_CDD, Per_VM
c. Predictors: (Constant), Per_NO_CDD, Per_VM, Per_JtW_WB
d. Predictors: (Constant), Per_NO_CDD, Per_VM, Per_JtW_WB, MATI_HH_K
e. Predictors: (Constant), Per_NO_CDD, Per_VM, Per_JtW_WB, MATI_HH_K, Per_Renter
f. Predictors: (Constant), Per_NO_CDD, Per_VM, Per_JtW_WB, MATI_HH_K, Per_Renter, Per_Unemp
g. Dependent Variable: Per_DC_Both

In terms of the MAUP effect, this shows how different geographical levels provide different significant variables. The excluded variables are shown in Table A1. Between the two models, variables that remain consistent are median after-tax income of households and visible minority in private households. The CSD model uses immigration as a significant variable however the NH model does not.

Table 4 shows the Pearson correlations between the variables chosen by the stepwise function in regards to the CSD geographic level. We are able to note that only median after tax income has a negative correlation with our study variable diabetes, meaning that when diabetes increases median income decreases and vice versa, which corresponds with other researches that lowest quintiles of household income have higher rates of diabetes than the upper quantiles, several studies have observed similar associations (Lord et al., 2020; Rabi at al., 2006). On the other hand, with the rate of immigration and visible minorities increases we can observe a positive correlation meaning that when diabetes increases, the rates do as well. Table 5 shows the coefficient values of the CSD model. In this table the coefficient for immigration is negative even though the correlation was positive. This happened because there is a high correlation between the immigration and visible minorities rate variables. It also shows that there is multicollinearity based on the VIF statistic, this due to the size of the CSD data and is spoken about more later on in the limitations section.

Table 4

CSD Final model Correlation values

		Per_DC_Both	MATI_HH_K	Per_VM	Per_Imm
Pearson Correlation	MATI_HH_K	-0.691	1.000	0.166	0.207
	Per_VM	0.186	0.166	1.000	0.968
	Per_Imm	0.082	0.207	0.968	1.000

Table 5

CSD Final model Coefficient values

Model		Unstandardized Coefficients		Standardized	t	Sig.	95.0% Confidence Interval for		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
3	(Constant)	18.284	0.947		19.317	0.000	16.358	20.210					
	MATI_HH_K	-0.079	0.012	-0.693	-6.422	0.000	-0.104	-0.054	-0.691	-0.745	-0.672	0.939	1.065
	Per_VM	0.116	0.037	1.327	3.147	0.003	0.041	0.191	0.186	0.480	0.329	0.061	16.268
	Per_Imm	-0.128	0.051	-1.060	-2.493	0.018	-0.232	-0.024	0.082	-0.398	-0.261	0.061	16.525

a. Dependent Variable: Per_DC_Both

Table 6 shows the Pearson correlation values between the variables chosen by the stepwise function in regards to the NH geographical level. Median income was a variable chosen in both models and can be noted that there was a negative correlation in both models as well. In this model, physical activity rate and renter rate have a negative correlation with diabetes. This makes sense as one would expect that someone who walks or bikes to work would not have any illness. It is interesting to note that in both models immigration rate had a positive correlation with diabetes but was only significant for the CSD model. Also immigration rate had a positive correlation with median income on the CSD level but negative on the NH model. The same thing occurs with the visible minority variable as it previously had a positive relationship with median income but it is negative on the NH level. This suggests that as the geographical extent of an area reduces then diabetes is affected in different ways by variables.

Table 6

NH Final model Correlation values

		Per_DC_Both	MATI_HH_K	Per_VM	Per_NO_CDD	Per_Unemp	Per_Renter	Per_JtW_WB
Pearson Correlation	Per_DC_Both	1.000	-0.405	0.423	0.700	0.534	-0.015	-0.467
	MATI_HH_K	-0.405	1.000	-0.178	-0.479	-0.610	-0.641	-0.245
	Per_VM	0.423	-0.178	1.000	0.078	0.499	0.287	0.014
	Per_Imm	0.379	-0.135	0.943	0.054	0.419	0.292	-0.040
	Per_NO_CDD	0.700	-0.479	0.078	1.000	0.404	0.041	-0.317
	Per_Unemp	0.534	-0.610	0.499	0.404	1.000	0.546	0.019
	Per_Renter	-0.015	-0.641	0.287	0.041	0.546	1.000	0.538
	Per_JtW_WB	-0.467	-0.245	0.014	-0.317	0.019	0.538	1.000

From the coefficient table for the NH model (Table 7) we can look at the coefficient beta value and see that there are three variables in the model that have negative relationships with diabetes. These variables are renter %, physical activity % and median after tax income. This validates the existing knowledge about predictors of diabetes, as research has shown that income, physical activity and a walkable neighbourhood environment can have a positive influence on diabetes prevention rates (India-Aldana et al., 2022; Lord et al., 2020; Rabi at al., 2006).

Table 7

NH Final model Coefficient values

Model		Unstandardized Coefficients		Standardized			95.0% Confidence Interval for		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
6	(Constant)	10.672	0.975		10.942	0.000	8.754	12.590					
	Per_NO_CDD	0.187	0.017	0.384	10.794	0.000	0.153	0.221	0.700	0.495	0.280	0.531	1.884
	Per_VM	0.042	0.004	0.326	10.615	0.000	0.034	0.050	0.423	0.489	0.275	0.712	1.405
	Per_JtW_WB	-0.100	0.012	-0.286	-8.104	0.000	-0.124	-0.075	-0.467	-0.393	-0.210	0.541	1.848
	MATI_HH_K	-0.049	0.008	-0.281	-6.408	0.000	-0.064	-0.034	-0.405	-0.320	-0.166	0.350	2.857
	Per_Renter	-0.043	0.008	-0.253	-5.704	0.000	-0.058	-0.028	-0.015	-0.288	-0.148	0.340	2.938
	Per_Unemp	0.300	0.066	0.188	4.574	0.000	0.171	0.429	0.534	0.235	0.119	0.396	2.526

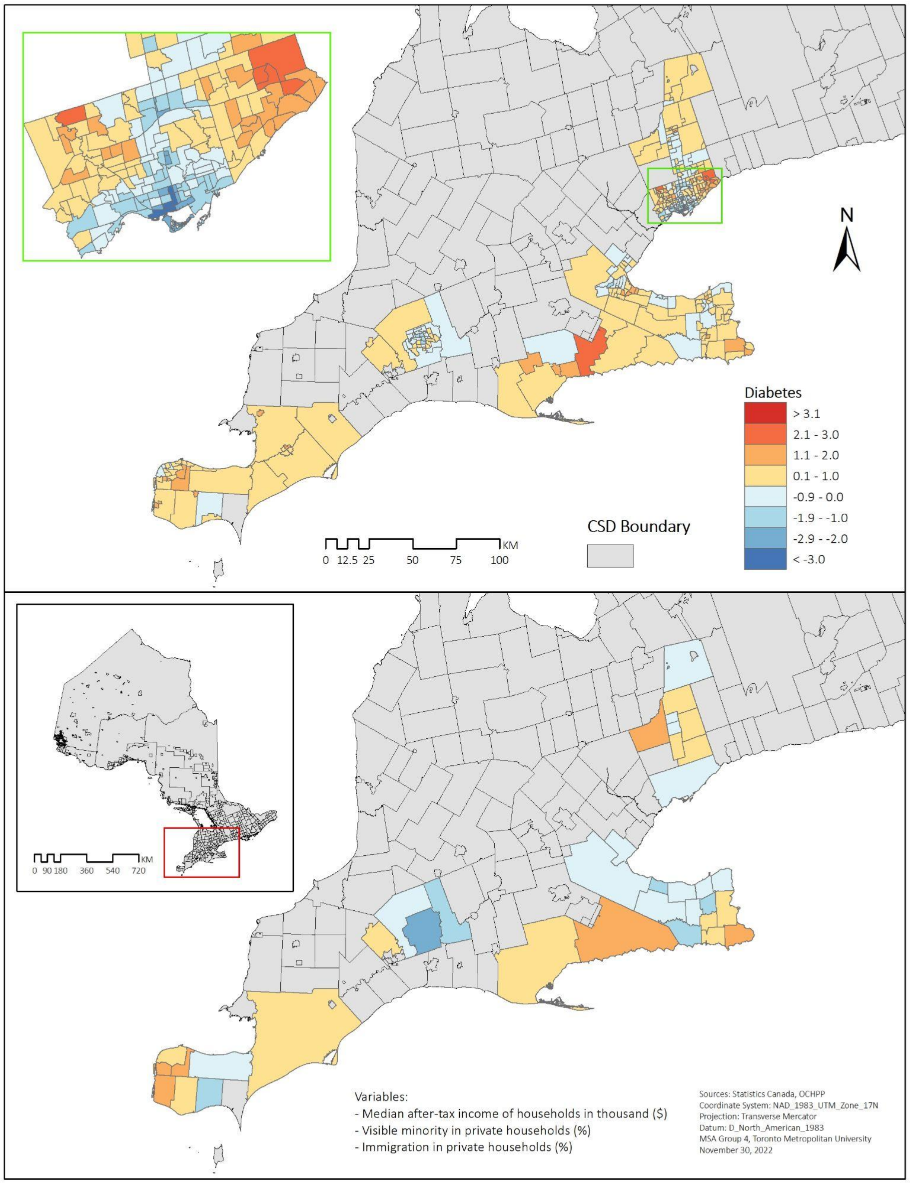
a. Dependent Variable: Per_DC_Both

The differences in the associations between immigration status and diabetes can be explained by the fact that the influence of immigration on diabetes varies depending on the country of origin. For instance, Kolpak and Wang's (2017) study about the prevalence of diabetes in Toronto determined that native born and North American ethnicity were associated with lower rate of diabetes, while foreign born, Asian place of birth has the opposite association with diabetes.

When examining the residual map (see Figure 6), it is important to note that the positive values on this map indicate that the model underpredicted diabetes prevalence and the negative values mean that the model overpredicted diabetes. When a residual has a value greater than three or less than negative three, one can call this residual an outlier. In order to compare the residuals on two different levels of data aggregation it was important to keep the variables the same between the two models. This meant using the variables from the stepwise CSD model in Table 2 to create a NH model which can be seen in Table A4. There are limitations to this approach as one variable in the new NH model is not significant. However this also helps us observe the residuals in relation to the MAUP effect and the problems it creates. On the NH level map we can observe a few outliers in the lower part of the downtown of Toronto. The model underpredicts many nieghborhoods near the east and west of Toronto and overpredicts nieghborhoods through the middle and lower end. The CSD map shows that there are not any outliers and the model does not overpredict or underpredict by much.

Figure 6

Map of Residual at the Neighbourhood and CSD levels



Predictive Analysis

Next, this study applies the regression coefficient results of different models to the neighbourhood and CSD levels to test the accuracy of the regression model in predicting other levels. Use the regression coefficient results at the CSD level to predict the neighbourhood and CSD level, and the regression coefficient at the neighbourhood level (using all variables) results to predict the CSD level. The regression coefficients and calculation formulas used in the three predictions are as follows:

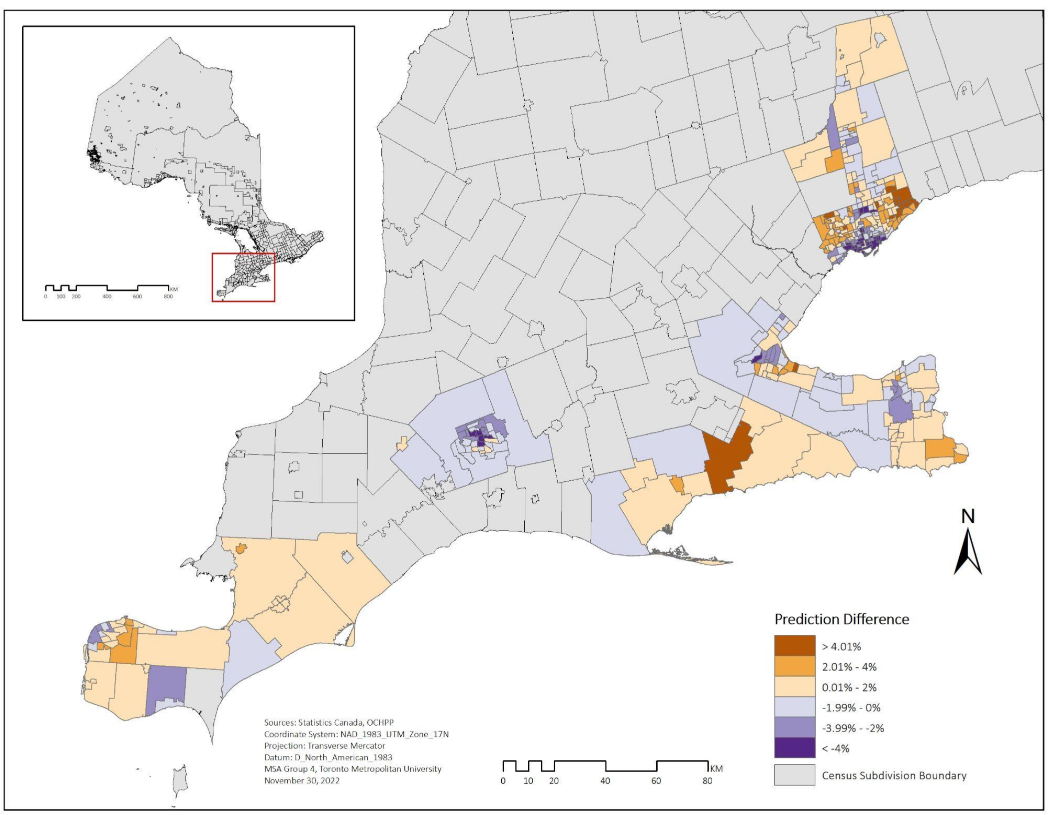
Predicted value (from CSD Model) = 18.284 + (-0.079*MATI_HH_K) + (0.116*Per_VM) + (-0.128*Per_Imm)

Predicted value (from Neighbourhood Model) = 10.672 + (0.187* Per_NO_CDD) + (0.042*Per_VM) + (-0.1*Per_JtW_WB) + (-0.049*MATI_HH_K) + (-0.043*Per_Renter) + (0.3*Per_Unemp)

After making predictions for different levels, this study subtracted the predicted value from the actual value, so that the error between the predicted and actual values could be obtained. It can be seen very clearly that when using the CSD-level regression coefficient to predict the neighbourhood level, the results show that there is a large gap between the actual value and the predicted value in many regions (see Figure 7). The model severely underestimates areas to the east of the City of Toronto, areas to the south of Hamilton Niagara Haldimand Brant; and severely overestimates areas to the central, south, and southwest of the City of Toronto, and areas to the west of Hamilton Niagara Haldimand Brant.

Figure 7

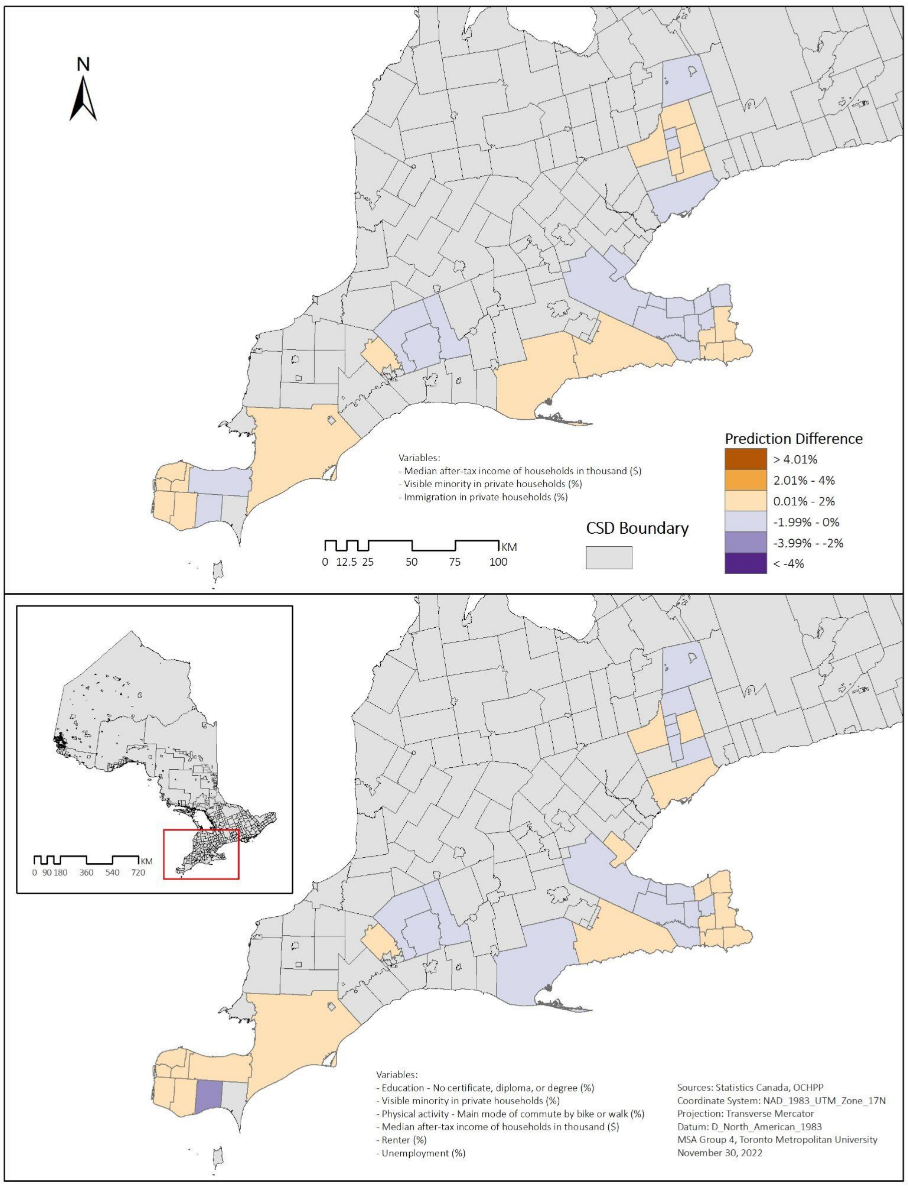
Map of Prediction Difference at the Neighbourhood levels



However, when the CSD level is predicted by the regression coefficient of the neighbourhood level, the situation becomes different (see Figure 8). The predictions of most regions are relatively accurate, much better than the prediction of neighbourhoods by CSD regression coefficient, and the prediction errors are significantly reduced.

Figure 8

Map of Prediction Difference at the CSD levels



When using 3 variables (MATI_HH_K, Per_VM, Per_Imm) for prediction, the error was reduced to between -1.98% and 1.87%. When using 6 variables (Per_NO_CDD, Per_VM, Per_JtW_WB, MATI_HH_K, Per_Renter and Per_Unemp) for prediction, the error was reduced to between -2.13% and 1.72%. This is much more accurate than the prediction error (-12.1% ~ 7.5%) obtained by predicting the neighbourhood through the CSD regression coefficient (see Table 10).

Table 10

Prediction Difference Statistics

	Prediction Difference		
	Neighbourhood with 3 Variables	**CSD with 3 Variables**	**CSD with 6 Variables**
Minimum	-12.104695	-1.980215	-2.128966
Maximum	7.495219	1.866078	1.716917
Mean	-0.402115	-0.015215	0.060871
Standard Deviation	2.885512	0.901523	0.876646

Conclusions

Overall, this study provided evidence that the MAUP does appear to influence results when using a linear regression model to predict diabetes prevalence between the two study areas (neighbourhoods and Census subdivisions). This study serves as a case study showing the importance of acknowledging the impact of the MAUP in local research. Namely, some significant predictors of diabetes varied across the two models, and the strength of prediction of diabetes by independent variables varied across the models. Further, other analyses, such as descriptive statistics and measures of spread, also showed variables differed across the models. Notably, one of our findings was inconsistent with our hypotheses and the literature, as the CSD model, of greater data aggregation, actually showed weaker predictions of diabetes. However, we attribute this to the sample size limitations at the CSD level (n = 37). Fortunately, the results of our regression were overall generally consistent with existing literature related to predictors of diabetes, posing less concern about the reliability of research in this area.

Limitations

In terms of limitations, it should be noted this study was restricted in some ways, which could be improved upon for future research. First, because we were unable to access raw data for diabetes prevalence across geographic units due to data restrictions and availability, we could not conclude patient characteristics. We had to create some of this data ourselves using what was available like census variables to explore the relationships between diabetes prevalence and SES. Access to raw data at different levels of aggregation would improve the power of these analyses and the ability to draw comparisons between scales. Second, this study used a smaller sample size, particularly problematic at the Census subdivision level where the sample only included 37 subdivisions. Because of this, the variables were not all normally distributed, posing problems for the reliability of analyses. Likewise, this creates problems of power, where effects are less able to be detected with smaller sample sizes. Consequently, this may also account for differences between the two models. To improve upon this, improving data availability would also enable using greater sample sizes for future research.

Recommendations

Despite these limitations, our study indicates the Ontario region is unsurprisingly not immune to the MAUP. This creates undoubtable challenges for researchers, however, we offer some recommendations. As such, we suggest using lower levels of data aggregation for health related data analysis, along with running analyses at multiple levels of aggregation to decrease a chance of missing important differences that are not evident at a larger scale or inflating effects that may not be so important. In addition, we recommend collecting and using individual data that has not been aggregated where possible, or to help strengthen findings, though challenging with privacy concerns. Finally we recommend using other exploratory and statistical analyses to reinforce the confidence in results before using these findings in health related decision making.

References

Agardh, E. E., Sidorchuk, A., Hallqvist, J., Ljung, R., Peterson, S., Moradi, T., & Allebeck, P. (2011). Burden of type 2 diabetes attributed to lower educational levels in Sweden. *Population Health Metrics, 9*, 60. https://doi.org/10.1186/1478-7954-9-60

Awuor, L., & Melles, S. (2019). The influence of environmental and health indicators on premature mortality: An empirical analysis of the City of Toronto's 140 neighborhoods. *Health & Place, 58,* 102155.

Cartone, A., & Postiglione, P. (2021). Principal component analysis for geographical data: The role of spatial effects in the definition of composite indicators. *Spatial Economic Analysis, 16*(2), 126-147. https://doi.org/10.1080/17421772.2020.1775876

Cressie, N. A. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems, 3* (2-3), 159-180.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning, 23*(7), 1025-1044.

Gariepy, G., Kaufman, J. S., Blair, A., Kestens, Y., & Schmitz, N. (2015). Place and health in diabetes: the neighbourhood environment and risk of depression in adults with Type 2 diabetes. Diabetic Medicine, 32(7), 944-950.

Government of Canada, S. C. (2022, January 24). *Leading causes of death, total population, by age group*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401

Hazell, E. C., & Rinner, C. (2020). The impact of spatial scale: Exploring urban butterfly abundance and richness patterns using multi-criteria decision analysis and principal component analysis. *International Journal of Geographical Information Science, 34*(8), 1648–1681.

India-Aldana, S., Kanchi, R., Adhikari, S., Lopez, P., Schwartz, M. D., Elbel, B. D., Rummo, P. E., Meeker, M. A., Lovasi, G. S., Siegel, K. R., Chen, Y., & Thorpe, L. E. (2022). Impact of land use and food environment on risk of type 2 diabetes: A national study of veterans, 2008–2018. *Environmental Research*, *212*, 113146. https://doi.org/10.1016/j.envres.2022.113146

Jones, K., Manley, D., Johnston, R., & Owen, D. (2018). Modelling residential segregation as unevenness and clustering: A multilevel modelling approach incorporating spatial dependence and tackling the MAUP. *Environment and Planning B: Urban Analytics and City Science, 45*(6), 1122–1141.

Kolpak, P., & Wang, L. (2017). Exploring the social and neighbourhood predictors of diabetes: a comparison between Toronto and Chicago. *Primary Health Care Research & Development, 18*(3), 291-299. https://doi.org/10.1017/S1463423617000044

Lee, D. W., & Rogers, M. (2019). Measuring geographic distribution for political research. *Political Analysis, 27*(3), 263-280.

Lord, J., Roberson, S., & Odoi, A. (2020). Investigation of geographic disparities of pre-diabetes and diabetes in Florida. *BMC Public Health*, *20*(1), 1226. https://doi.org/10.1186/s12889-020-09311-2

Mitra, R., & Buliung, R. N. (2012). Built environment correlates of active school transportation: Neighborhood and the modifiable areal unit problem. *Journal of Transport Geography*, *20*(1), 51–61.

Moat, J., Bachman, S. P., Field, R., & Boyd, D. S. (2018). Refining area of occupancy to address the modifiable areal unit problem in ecology and conservation. *Conservation Biology, 32*(6), 1278-1289. https://doi.org/10.1111/cobi.13139

O'Campo, P., Wheaton, B., Nisenbaum, R., Glazier, R. H., Dunn, J. R., & Chambers, C. (2015). The Neighbourhood Effects on Health and Well-being (NEHW) study. *Health & Place, 31*, 65–74.

Ontario Community Health Profiles Partnership. (2022). www.ontariohealthprofiles.ca

Openshaw, S., Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem Statistical Application in the Spatial Sciences. London: Pion. . (1979). A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, 127-144.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. *CATMOG, 38.* Norwich, England: Geobooks.

Piccolo, R. S., Duncan, D. T., Pearce, N., & McKinlay, J. B. (2015). The role of neighborhood characteristics in racial/ethnic disparities in type 2 diabetes: Results from the boston area community health (BACH) survey. *Social Science & Medicine (1982), 130*, 79-90. https://doi.org/10.1016/j.socscimed.2015.01.041

Quick, M., & Revington, N. (2022). Exploring the global and local patterns of income segregation in Toronto, Canada: A multilevel multigroup modeling approach. *Environment and Planning B: Urban Analytics and City Science, 49*(2), 637–653.

Rabi, D. M., Edwards, A. L., Southern, D. A., Svenson, L. W., Sargious, P. M., Norton, P., ... & Ghali, W. A. (2006). Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. *BMC Health Services Research, 6*(1), 1-7.

Saib, M., Caudeville, J., Carre, F., Ganry, O., Trugeon, A., & Cicolella, A. (2014). Spatial relationship quantification between environmental, socioeconomic and health data at different geographic levels. *International Journal of Environmental Research and Public Health, 11*(4), 3765-3786. https://doi.org/10.3390/ijerph110403765

Statistics Canada. (2022). Table 13-10-0096-07 Diabetes, by age group. https://www150-statcan-gc-ca.ezproxy.lib.ryerson.ca/t1/tbl1/en/tv.action?pid=1310009607

Tompkins, J. W., Luginaah, I. N., Booth, G. L., & Harris, S. B. (2010). The geography of diabetes in London, Canada: The need for local level policy for prevention and management. *International Journal of Environmental Research and Public Health, 7*(5), 2407-2422. https://doi.org/10.3390/ijerph7052407

Vallée, J., Shareck, M., Le Roux, G., Kestens, Y., & Frohlich, K. L. (2020). Is accessibility in the eye of the beholder? Social inequalities in spatial accessibility to health-related resources in Montréal, Canada. *Social Science & Medicine*, 245, 112702.

Wang, Y., & Di, Q. (2020). Modifiable areal unit problem and environmental factors of COVID-19 outbreak. *The Science of the Total Environment, 740*, 139984-139984. https://doi.org/10.1016/j.scitotenv.2020.139984

Xu, P., Huang, H., & Dong, N. (2018). The modifiable areal unit problem in traffic safety: Basic issue, potential solutions and future research. *Journal of Traffic and Transportation Engineering, 5*(1), 73-82.

Appendix A

Table A1

Excluded Variables from NH model

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	VIF	Minimum
1	MATI_HH_K	-0.09	-2.132	0.034	-0.111	0.770	1.298	0.770
	Per_VM	0.371	11.550	0.000	0.518	0.994	1.006	0.994
	Per_Imm	0.342	10.374	0.000	0.478	0.997	1.003	0.997
	Per_Unemp	0.3	7.936	0.000	0.385	0.836	1.196	0.836
	Per_Renter	-0.043	-1.162	0.246	-0.061	0.998	1.002	0.998
	Per_JtW_WB	-0.272	-7.387	0.000	-0.361	0.900	1.112	0.900
2	MATI_HH_K	-0.023	-0.626	0.532	-0.033	0.750	1.333	0.750
	Per_Imm	-0.07	-0.723	0.470	-0.038	0.110	9.097	0.110
	Per_Unemp	0.126	3.114	0.002	0.162	0.616	1.623	0.616
	Per_Renter	-0.162	-5.009	0.000	-0.255	0.917	1.090	0.913
	Per_JtW_WB	-0.289	-9.536	0.000	-0.448	0.898	1.113	0.893
3	MATI_HH_K	-0.225	-6.282	0.000	-0.314	0.579	1.726	0.569
	Per_Imm	-0.235	-2.685	0.008	-0.140	0.106	9.435	0.106
	Per_Unemp	0.192	5.358	0.000	0.271	0.598	1.673	0.598
	Per_Renter	0.012	0.317	0.752	0.017	0.593	1.687	0.580
4	Per_Imm	-0.239	-2.878	0.004	-0.150	0.106	9.436	0.105
	Per_Unemp	0.108	2.671	0.008	0.139	0.449	2.229	0.435
	Per_Renter	-0.184	-4.284	0.000	-0.220	0.386	2.591	0.377
5	Per_Imm	-0.154	-1.812	0.071	-0.095	0.098	10.212	0.098
	Per_Unemp	0.188	4.574	0.000	0.235	0.396	2.526	0.340
6	Per_Imm	-0.046	-0.525	0.600	-0.028	0.089	11.193	0.084

Table A2

Excluded Variables from CSD model

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	VIF	Minimum
1	Per_NO_CDD	0.197	1.079	0.288	0.182	0.447	2.236	0.447
	Per_VM	0.309	2.716	0.010	0.422	0.972	1.028	0.972
	Per_Imm	0.235	1.956	0.059	0.318	0.957	1.045	0.957
	Per_Unemp	0.354	1.958	0.058	0.318	0.423	2.365	0.423
	Per_Renter	-0.049	-0.280	0.782	-0.048	0.508	1.968	0.508
	Per_JtW_WB	-0.206	-1.379	0.177	-0.230	0.649	1.540	0.649
2	Per_NO_CDD	0.354	2.127	0.041	0.347	0.411	2.431	0.411
	Per_Imm	-1.06	-2.493	0.018	-0.398	0.061	16.525	0.061
	Per_Unemp	0.112	0.510	0.614	0.088	0.269	3.712	0.269
	Per_Renter	-0.361	-2.063	0.047	-0.338	0.376	2.661	0.376
	Per_JtW_WB	-0.327	-2.424	0.021	-0.389	0.605	1.653	0.595
3	Per_NO_CDD	0.238	1.381	0.177	0.237	0.357	2.799	0.053
	Per_Unemp	0.003	0.012	0.990	0.002	0.257	3.896	0.053
	Per_Renter	-0.285	-1.675	0.104	-0.284	0.359	2.784	0.058
	Per_JtW_WB	-0.223	-1.519	0.139	-0.259	0.488	2.050	0.049

Table A3

NH model with 3 variables

		Unstandardized Coefficients		Standardized			95.0% Confidence Interval for		Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	14.698	0.656		22.406	0.000	13.408	15.988					
	MATI_HH_K	-0.059	0.008	-0.338	-7.485	0.000	-0.074	-0.043	-0.405	-0.366	-0.331	0.958	1.043
	Per_VM	0.057	0.017	0.443	3.289	0.001	0.023	0.091	0.423	0.170	0.145	0.108	9.287
	Per_Imm	-0.016	0.025	-0.085	-0.634	0.527	-0.065	0.033	0.379	-0.033	-0.028	0.109	9.160

a. Dependent Variable: Per_DC_Both

Table A4

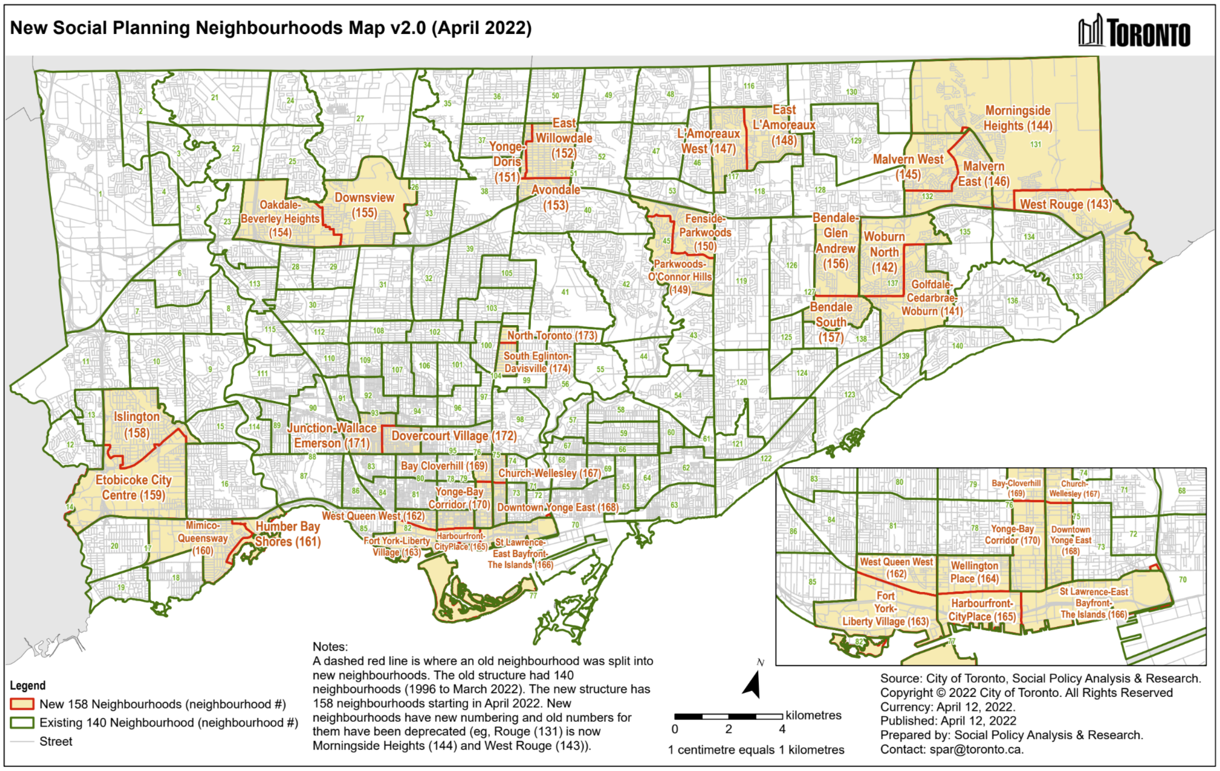
Coefficients for All NH models

		Unstandardized Coefficients		Standardized
Model		B	Std. Error	Beta
1	(Constant)	6.200	0.335	
	Per_NO_CDD	0.341	0.018	0.700
2	(Constant)	4.893	0.309	
	Per_NO_CDD	0.327	0.016	0.672
	Per_VM	0.048	0.004	0.371
3	(Constant)	6.359	0.316	
	Per_NO_CDD	0.282	0.015	0.579
	Per_VM	0.049	0.004	0.382
	Per_JtW_WB	-0.101	0.011	-0.289
4	(Constant)	10.519	0.727	
	Per_NO_CDD	0.216	0.018	0.443
	Per_VM	0.046	0.004	0.354
	Per_JtW_WB	-0.135	0.011	-0.387
	MATI_HH_K	-0.039	0.006	-0.225
5	(Constant)	12.772	0.884	
	Per_NO_CDD	0.202	0.017	0.415
	Per_VM	0.050	0.004	0.389
	Per_JtW_WB	-0.113	0.012	-0.324
	MATI_HH_K	-0.058	0.008	-0.335
	Per_Renter	-0.031	0.007	-0.184
6	(Constant)	10.672	0.975	
	Per_NO_CDD	0.187	0.017	0.384
	Per_VM	0.042	0.004	0.326
	Per_JtW_WB	-0.100	0.012	-0.286
	MATI_HH_K	-0.049	0.008	-0.281
	Per_Renter	-0.043	0.008	-0.253
	Per_Unemp	0.300	0.066	0.188

Appendix B

Figure B1

New Neighbourhood Boundaries in Toronto



Note: Taken from the City of Toronto, Social Policy Analysis & Research.