Use of a dwelling-referenced geographic information system to characterize urban tuberculosis

Ian Wanyeki
Department of Epidemiology and Biostatistics
McGill University, Montreal

September 2003

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Masters of Science, Epidemiology

© Ian Wanyeki 2003.

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-98837-6
Our file Notre référence
ISBN: 0-612-98837-6

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Acknowledgements

I am extremely grateful to both my supervisors as well as my supervisory committee members for their guidance and useful suggestions as I have prepared this thesis.

I am also grateful to the Montreal Chest Institute Research Centre and the Canadian Institutes of Health Research for providing me with some financial assistance. The city of Montreal provided a crucial database and data management tool for university use. This study was carried out in collaboration with the Respiratory Epidemiology Unit, the McGill Department of Geography, and la Direction de Santé Publique de Montréal-Centre.

I would also like to thank my wife for her continued support and patience as I undertook this project.

Table of Contents

Contents **Page**

Glossary

CLSC	Centre local de services communautaires (Community health clinic)
GIS	Geographic information system
SIG	Système d'information géographique
GPS	Global positioning systems
MADO	Maladies à déclaration obligatoire (Reportable diseases)

List of tables   Page

List of figures

Page

List of appendices

Page

Abstract

Although urban overcrowding is frequently cited as promoting spread of TB, it is difficult to separate the influence of housing from that of other socio-economic and demographic factors. While poor housing conditions of the early 20^{th} century were linked to contagion, links between dwelling features and TB are less clear with current housing standards. Access to specific dwelling attributes is not usually an option and most previous research on TB has relied on socio-economic variables available from aggregated census data. This thesis evaluates the potential role of dwelling-specific attributes in describing TB occurrence in an urban setting.

Using ArcView 3.2 software, all active TB cases reported in the former city of Montreal 1996-2000 were precisely geo-coded, that is mapped to the corresponding residential address. For comparison, using a case-cohort approach control dwellings were randomly selected from the municipal dwelling GIS, with a 10:1 ratio. We identified 595 case and 5950 control dwellings. Census tract data from the 1996 Canadian Census as well as dwelling characteristics from the Montreal housing database were attributed to both case and control dwellings. Multivariate logistic regression was used with dwelling status (case vs. control) as the dependant variable, to evaluate the independent influence of crowding and other socio-demographic factors.

A high-precision housing GIS complemented census data in pinpointing and characterising the occurrence of TB in Montreal. It provided a more refined assessment of the impact of local crowding, after adjustment for other important factors. Several housing risk factors were associated with TB occurrence independently of key socio-demographic predictors. These included higher dwelling density, younger dwelling age, and lower dwelling value and higher number of stories in the building. However local crowding (floor area per person) was not found to be a significant risk factor.

Résume

Bien que les milieux urbains surpeuplés soient considérés comme un facteur favorisant la transmission de la TB, il est difficile de différencier l'influence du logement lui-même des autres facteurs socioéconomiques et démographiques. Même si les conditions d'insalubrité au début du 20^{ieme} siècle étaient liées à la transmission, le lien entre les caractéristiques du bâtiment et la TB est présentement moins évident. Souvent les caractéristiques des logements ne sont pas disponibles et la plupart des recherches utilisent les données socioéconomiques du recensement. Cette thèse examine le rôle potentiel des caractéristiques du logement dans l'occurrence de la TB en milieu urbain.

Tous les cas actifs de TB, entre 1996 et 2000 dans la ville de Montreal, ont été précisément cartographiés. Les contrôles ont été sélectionnés au hasard pour un ratio de 10 contrôles par 1 cas. 595 cas et 5950 contrôles était identifié. Les données du recensement Canadien de 1996 et du SIG ont été assignées aux logements du cas et des contrôles. La régression logistique multivariée a été utilisée avec le status du logement (cas versus contrôle) comme variable dépendante pour évaluer l'effet indépendant du surpeuplement et des autres facteurs sociodémographiques.

Le SIG à haute précision a été une source des données complémentaires dans la caractérisation et la cartographie de l'occurrence de la TB à Montreal. Il aide dans l'évaluation de l'effet du surpeuplement local après avoir considéré les autres facteurs d'importance. Plusieurs caractérisriques du logement étaient associées à l'occurrence de la TB indépendamment des facteurs sociodémographiques clés : une densité élevée, la nouveauté du logement, une valeur bas du logement ainsi que un nombre élevé des étage dans la bâtisse. Néanmoins, le surpeuplement local (l'aire de plancher par personne) n'a pas été trouvé comme étant un facteur de risque significatif.

Introduction

Tuberculosis (TB) continues to be a major issue in global public health. Despite the availability of medicines not available 50 years ago, deaths are increasing in both Eastern Europe and Africa, after almost 40 years of decline (Tuberculosis Fact sheet, 2002). The breakdown in health services, the spread of HIV/AIDS and the emergence of multi-drug resistant TB are factors contributing to the worsening impact of this disease. The number of refugees and displaced people in the world is also increasing due to unequal opportunities and warfare. TB spreads quickly in crowded refugee camps and shelters. It is difficult to treat mobile populations, as treatment takes at least six months and should ideally be closely supervised. Emigration from countries with high prevalence contributes to the burden in Canada, where TB is increasingly concentrated in the larger cities (Long, 1999). In large urban centres such as Montreal, that receive a large proportion of the immigrants to Canada (approximately 70%) it is therefore essential to identify high risk neighborhoods and promote access to health care among new migrants and their neighborhoods.

The incidence of TB in Montreal varies widely by neighborhood (Rivest, 1999). Much of this variation reflects the socio-demographic characteristics of inhabitants and their general health, but it may also reflect additional conditions that promote local spread of the disease. Montreal is distinct among Canadian cities due to a high level of residential segregation based on household income (Ross, 2001). Many neighborhoods are therefore very homogeneous in terms of affluence or poverty (figure 2). Several recent studies report an inverse relationship between socio-economic status and health (World development report, 1993), and the incidence of TB in particular correlates with poverty in various settings (Mangtani, 1995, Elender, 1998, Dennis, 2000). It is therefore important to investigate both specific locality features which include neighborhood and housing characteristics as well as socio-economic characteristics in epidemiological research in order to try to understand the separate influence of housing and neighborhood from that of socio-economic and demographic variables that are associated with TB.

General objective of the study

Our main goal was to assess the value added by using a high precision geodatabase (geographic database) to improving our understanding of local urban TB epidemiology. By exploiting the City housing Geographic information system (GIS)[1], containing detailed City of Montreal[2] housing data and linking this to census data and public health surveillance data, incident TB cases that occurred between 1996-2000 were mapped to their corresponding residential dwelling, which was then identified as a case dwelling and with the unit of analysis being a dwelling. Properties of the case dwellings were then compared to a group of randomly distributed control dwellings.

The City housing GIS enabled us to map, identify and analyze features of houses and neighborhoods that may tend to collect households at high risk of TB. Combined use of both the City housing GIS and census data allowed us to investigate local conditions (possible exposures) that may be important in promoting the transmission of TB. These included specific housing conditions such as crowding and housing quality. Additionally, we investigated socio-demographic factors such as income and foreign birth that relate more to neighborhood. This enabled us to evaluate the relationship between TB incidence and poverty while adjusting for foreign birth and specific housing characteristics. Census data allowed us to obtain socio-economic variables at the level of the census tract. The census was the sole data source readily available that allowed us to estimate the residential distributions of immigrant populations by their country of birth.

Specific objectives necessary to this project include (1) the precise mapping of all incident TB cases on the Island of Montreal for the period 1996 to 2000; (2) characterisation of variations in the occurrence of the disease according to attributes of individual dwellings; (3) creation of a crowding index (floor area per person) that can

---

[1] Terminology For the purposes of this thesis the City of Montreal Housing database also called the SIURS (Système d'information urbaine à reference spaciale) will be referred to as the City housing GIS.
[2] Until January 2002 the Island of Montreal consisted of 28 suburban municipalities, as well as the municipality of Montreal which I continue to refer to as the former City of Montreal in this thesis.

be applied at a fine level of analysis, namely the city block; and (4) within the former city of Montreal model the associations between TB and variables which distinguish poverty, crowding and birthplace.

The key hypotheses are centred on the issues of crowding: (1) Occurrence of TB is associated with residential crowding as defined by floor area per person at the city block level, even after adjustment for foreign birth; (2) the highest levels of crowding play a role in maintaining higher levels of TB within certain neighbourhoods; (3) poorer neighbourhoods are also expected to be at higher risk as indicated via proxy variables and (4) other properties of the housing stock, (notably dwelling value) an indicator of socio-economic status will be associated with increased risk.

Background and literature review

A) About TB

TB is a contagious disease that is caused by *Mycobacterium tuberculosis* and is transmitted through the air. Only people who are sick with active pulmonary or laryngeal TB are infectious (Reichman et al., 2000). When infectious people cough, sneeze, talk or spit, bacilli are propelled into the air thus causing spread via airborne transmission of the infectious droplets. This is a very important mode of transmission for communicable diseases such as TB which can be spread rapidly from person to person via an airborne route (Long et al., 1999). Some environmental conditions which promote transmission are well known from classic experimental research poor ventilation and crowding, which increase exposure to infectious persons and the droplets they produce (Riley et al., 1962).

B) TB globally

In 1997, the World Health Organization (WHO) estimated that 32% of the global population was infected with *Mycobacterium tuberculosis*. They estimated that between 2002 and 2020, approximately one billion people will be newly infected, over 150 million people will get sick, and 36 million will die of TB if control is not further strengthened (TB Fact sheet, 2002). With the spread of the human immunodeficiency virus (HIV) recharging the global resurgence of tuberculosis and the growing emergence of multi-drug resistant strains (Long, 2000) concerted efforts will be required on a global scale to manage this epidemic.

C) TB in Canada

Tuberculosis was a major cause of morbidity and mortality in Canada throughout the first half of the 20th century. Thanks to improvements in general living conditions, public health measures to interrupt transmission and the advent of antibiotic therapy, Canadian TB disease and death rates declined rapidly after the mid 1940s (Figure 1) (Njoo, 1998). However, following several decades of continuous decline the notification rate has leveled off since 1987 to the current level of slightly less than 6 per 100,000 population, which corresponds to approximately 2,000 cases per year (Long, 2000).

TB transmission has historically been associated with population movements and epidemiologists as well as geographers have long been aware of the close links between spread of disease and human mobility (Kamel, 1997, Wessen, 1974). More recently globalization, internal conflicts and war have caused the displacement of large numbers of people who are often from countries with a high incidence of TB, and who also may be in crowded facilities such as refugee camps before departure. This increases the potential for the rapid spread of and changes in the spatial patterns of disease prevalence (Kazmi et al., 2001) and is an important issue that has implications on the burden of TB globally and in particular within developed countries such as Canada that receive a large number of immigrants on a yearly basis. Migrants can act both as receivers of new disease as well as transmitters (Prothero, 1977).

Recently the epidemiology of tuberculosis in Canada has changed, with major shifts in the origins of reported cases and their locations (Long, 2000). Examination of its distribution in 1998 reaffirms the trend that the disease has retreated into geographically and demographically distinct groups (Long et al., 1999).

In 1998, the proportion of Canadian-born non-Aboriginal cases had decreased to 19% from 48% in 1981. Recent immigrants made up the majority of cases (64% as compared to 37% in 1981), while the proportion of cases from First Nations people (Status and Non-Status Indians, Métis and Inuit) remained unchanged at 15% (Long, 2000). Recent immigrants constitute only approximately 15% of the Canadian population, and their share of TB is four times that expected for their number in the population (Fanning, 1998). This increase is consistent with the shift in immigration patterns over the last 30-40 years (Long, 2000) as the majority of immigrants and refugees now come from high incidence countries.

D) Geographic distribution of TB in Canada

Within Canada, 75 % of the TB cases in 1995 occurred in the provinces of Ontario, Québec and British Colombia, where a disproportionately large number of cases were reported from their large urban centres Toronto, Montreal and Vancouver. In the province of Québec where incidence has decreased steadily over time to a low of 5 / 100,000 in 1995 the incidence on the Island of Montreal plateaued at 11 in the early 1980's and has remained at this level through the 1990's (Rivest et al., 1998). In several other developed countries TB has increasingly become an urban disease. In the United States in 1998, 75% of all TB cases were reported from metropolitan areas of 500,000 persons or more (Long, 2000). By 2000, cases of TB in London, England had risen to account for approximately 50% of the national burden (Iskander, 2001).

There is a correlation between TB occurrence in large urban cities and the proportion of immigrants found in such cities. This may be due to the fact that these same cities also attract a large number of immigrants from high prevalence countries and this contributes to the maintenance of higher levels of incidence within these urban centres. It may also be due to a relatively large presence of high risk groups such as IV drug users, the homeless and those with HIV.

The Island of Montreal accounts for more than half of the reported TB cases in Québec (Rivest et al., 1998). Mean annual incidence among immigrants (37.5/100,000) between 1992 - 1995 was ten times that of Canadian-born (3.3/100,000) (Rivest et al., 1998). The incidence of TB on the Island of Montreal varies widely by district. TB incidence was found to range from 3.1 to 40 /100,000 when stratified by community clinic (CLSC) catchment area (Rivest et al., 1999). This variance appears to correlate with Montreal's distinct neighborhoods (figure 2), which are relatively homogeneous in terms of affluence and poverty (Ross et al., 2001). The lower incidence rates were all seen in wealthy and middle class suburbs while the high incidence rates appeared in poorer neighborhoods that also are home to large immigrant populations. This phenomenon has been observed in several other cities, where people living in lower income neighborhoods were at highest risk (Long et al. 1999, Barr et al., 2001).

E) TB management and control in Montreal

Throughout Canada the reporting of active TB is required by legislation. Notifications of active TB cases are sent to the island wide Montreal department of public health, which has a centralized tuberculosis unit, responsible for the data. All reported cases are investigated by a public health nurse who obtains clinical and epidemiological information, ensures that appropriate treatment is initiated, and that those in close contact with the patient are tested and referred for further assessment if required.

Cases are classified as either culture confirmed or clinically confirmed. If *Mycobacterium tuberculosis* were cultured from sputum, body fluids or tissue, it is classified as culture confirmed. Cases are classified as clinically confirmed if the signs and symptoms are clinically compatible with TB, if there is radiological and pathological evidence of active disease, and if treatment is prescribed by a physician (Rivest et al., 1998).

All isolates are routinely tested for susceptibility to the first-line TB antibiotics. Confirmation and drug susceptibility testing is carried out by the provincial reference laboratory (Laboratoire de santé publique du Québec, LSPQ) according to a standard protocol. The results of the drug susceptibility tests are sent to the Montreal public health department and entered into the province-wide Maladies à déclaration obligatoire (MADO) database, which systematically records data on all notifiable diseases.

This comprehensive management and follow up aims at providing complete case ascertainment and aids in the management of TB within the province of Québec.

F) TB and poverty

i) About poverty

Most health indicators confirm the association between ill heath and poor socio-economic status (Spence, 1993), and TB is no exception. The tendency for TB to occur within poor neighborhoods (Lemer, 1993) has long been recognized by public health officials, and TB rates continue to correlate with poverty in several settings (Mangtani et al., 1995, Lemer, 1993). A study carried out in Liverpool found that the historical association between TB and poverty holds true even at present (Spence, 1993).

Poverty is multifaceted and complex and on its own cannot explain poor health. Moreover, levels of poverty are often relative to the wealth of society as a whole (Spence, 1993). It is often intrinsically linked to several other social aspects, which include overcrowding, poor nutrition, a weaker immune system, unemployment, smoking, immigration, homelessness, and drug abuse. The increased risk of TB among minorities is probably due to confounding by several of these social and economic factors that increase risk of exposure to TB and development of active disease once infected, rather than to genetics as shown by a study carried out in the US by the CDC (Cantwell et al., 1998). Several factors that help explain the relationship between TB and poverty in urban centres within a North American context are discussed below.

Poor nutrition, which compromises one's immune system makes a person more susceptible to attack by an infecting bacterium (McMurray et al., 1986, Strachan et al., 1995). When populations are under severe economic pressure or are perhaps homeless, drug users, or unemployed their diet and a healthy living routine are often compromised, which thus may result in a weakened immune system.

Infection by the HIV virus also weakens the immune system, leaving the body open to attack by opportunistic microorganisms such as TB. Worldwide, TB is the most common immediate cause of death among HIV-infected individuals and accounts for almost one-third of AIDS deaths annually. Those at highest risk of HIV include IV drug users, the homeless and people from countries with high incident rates of HIV. In Canada, there has been limited data on the prevalence of M. tuberculosis-HIV co-infection.

At the extreme of poverty are the homeless, who often suffer from a wide range of medical problems (Hwang, 2001). Numerous studies have indicated that factors which include living on the street, living in various institutional settings, HIV infection, and drug use, are all responsible for a large proportion of the TB cases reported annually (Rusen et al., 1999, El Sahly et al., 2001).

ii) Overcrowding

The role that crowding plays in TB transmission is complex as crowding varies from place to place and crowding within the western world takes on different meanings from crowding in a developing country or in a refugee camp. Additionally, the definition and the degree of residential overcrowding vary from study to study as well as from location to location. Nevertheless, residential overcrowding may be an important factor in the spread of TB (Acevedo-Garcia, 2001, Riley, 1962) attributable to the infectious nature of TB and its ability to spread via the airborne route.

Measures of persons per room and persons per thousand m^2 of floor area are conventionally referred to as crowding and might be pertinent to issues of ventilation of the dwelling and transmission within the family as well as transmission within buildings or other small areas. Measures of households per km^2 and persons per km^2 are conventionally referred to by planners as measures of residential density which is not the same as crowding. They are likely to have a very strong relationship to contagion, but this is fairly complex, as it might be associated with use of mass transit and all kinds of other inner-city characteristics.

In urban centers within Canada, there may be dwellings that are extremely overcrowded (10 people and more living within a two bedroom dwelling). Such extreme instances of overcrowding are likely to be due to lack of affordable housing and would probably play a major role in aiding the transmission of TB. Those who may live in such situations are likely to be recent immigrants who are from high incidence TB countries and whose socio-economic status tends to be low, thus making crowding a complex factor to study in isolation. Crowding remains however, an important risk factor as it is generally linked to other important TB risk factors. A particular problem of studying crowding is the fact that the extreme cases of 'crowding' are not well reflected in aggregated data. The median persons/dwelling in any census tract in Montreal is 2.1 with an inter-quartile range of 1.9 to 2.3. Crowding cannot be appropriately assessed without household-level information.

A study carried out by the US Centers for disease control indicated that there was a threshold value for crowding and the increased risk of developing TB was significantly higher once this crowding threshold was reached. Their measure of crowding was defined as median number of persons per room per household. They divided several socio-economic variables into quartiles and evaluated the association between the socio-economic indicators and the probability of being a TB case. They observed that the risk of TB increased with lower socio-economic status for several socio-economic variables studied, with crowding having the greatest impact (Cantwell et al., 1998).

A population of low purchasing power has the least choice of housing, but the types it is offered depend on the history of the particular city. One tends to assume, for example, that this will be older housing, rental housing, and under maintained or poorly heated housing. We shall see, however, that the Montreal housing stock is rather distinctive.

iii) Immigration

Due to several reasons that include settling in familiar cultural surroundings and cost of living, immigrants often settle in areas of the City that already contain a higher percentage of recent immigrants. It has been argued that since these areas are often poorer and more densely populated, with the incidence of TB already elevated, the likelihood of contracting the disease increases.

Among recent immigrants, the risk of developing active TB varies according to region of origin, with particularly high rates seen among immigrants from countries shown in figures 3 and 4. Montreal, which attracts a lot of immigrants from Francophone and former Francophone speaking countries, has a large number of cases among the Haitian born and Vietnamese born populations (Rivest et al., 1998) (figure 3).

In Britain, it was observed that that the risk of developing TB was highest among recent immigrants[3], with rates decreasing fairly rapidly after the first few years. Even many years after arrival, incidence was higher among immigrants as compared to the local population (McCarthy, 1984). This was also observed in the US (Zuber et al., 1997).

It is speculated that due to the negative stigma attached to those who have TB, they may be reluctant to seek treatment. The delays in seeking treatment are a major cause of ongoing transmission. In addition, access to available health care centres, or inadequate knowledge of the available treatment options, may be another hindrance. One's legal status or lack of status may cause people not to seek treatment due to fear that they may be deported. These are often described as barriers to care, and they must be considered in order to understand why higher levels of TB are maintained in areas with a high proportion of immigrants.

[3] Recent immigrants include people who immigrated between 1991 and 1996.

G) Geographic information systems as an investigation tool

All the life-style and environmental factors that affect human health are associated with where people live, and for this reason residential locations offer a valuable source for epidemiological research studies on health and the environment. Over a century ago scientists began to employ the potential of maps for understanding the spatial dynamics of disease (de Savigny et al., 1995). One of the most cited examples is John Snow's work showing that cholera spreads through infected water, a finding he was able to demonstrate with the aid of crude paper maps (Snow, 2002). GIS enables electronic plotting and provides a basis for more elaborate analysis as explained below.

Essentially, a GIS is a computer-assisted information management system for geographically referenced data. It contains two closely integrated databases one spatial (location) and the other statistical (attribute) (de Savigny et al., 1995).

The spatial database identifies location in terms of digital coordinates, usually derived from paper maps, from global positioning systems (GPS), or satellite imagery. These may be points (for example, the centroid of a health clinic or house), lines (rivers and roads), or polygons (district health units). The attribute database contains information about the characteristics or qualities associated with the location, for example, well depth, population in the district, immunization coverage, number of health personnel at the clinic, or the type of road access. Figure 5 shows a simple example of how layers of information can be overlaid. Each layer contains an underlying table with associated attributes.

As an example from Figure 5, the Montreal health districts layer would outline the health districts (CLSC), coupled with population and disease rates for each district, while the census tract layer would identify the tracts coupled with census variables (socio-economic and demographic)

Traditional database software and statistics programs are appropriate for analyses of attributes belonging to a single set of entities or, with more effort perfectly nested sets of entities, whereas the analytical tools available within a GIS make it possible to "overlay" data for different entities. This is similar to laying transparent maps for each database on top of each other (Figure 5). One of the main assets of a well developed GIS is the ease with which different layers of information can be combined by joining based on spatial location. As an example using figure 5, if one knows the location of a case it can immediately be linked to the correct census tract and health district. This kind of analysis may also be possible via traditional methods but it would be cumbersome, would require a lot of effort, many more steps and be more prone to error.

GIS includes hardware and software to enable the integration of various layers of data. It allows data to be viewed and analyzed spatially (Vine et al., 1998). It contains software for mapping and has an embedded relational database component that makes the management of data very organized (Stone, 2001). The relational database component enables the linkage of nongeographic attributes or geographically referenced data at various levels of scale to a graphic map, thus allowing for a wide array of data processing and display. It can be used to input, store, manage, analyze and display data.

Information systems based on GIS technology are used for many purposes ranging from assisting public health departments with disease control efforts to modeling and forecasting for early warning systems. They are also useful in explaining disease patterns in relation to socio-demographic and natural environments (Tim, 1995). They are often used in combination with a statistical modeling technique such as logistic regression for the study of infectious diseases (Moore et al., 1999). A GIS can be used to investigate questions about location (what are the attributes at a specific place?), condition (where are the sites that possess certain attributes?), trends (how do attributes change spatially over time?), routing (what is the shortest/least expensive/most cost-effective path between places?), and patterns (what is the distribution of attributes and the process/reason accounting for their distribution?). A GIS can also be used to simulate "what if" scenarios (de Savigny et al. 1995).

In epidemiological studies, GIS is often used to map disease and risk factors. This mapping of cases, as well as environmental factors that may be associated with disease vectors, aid in surveillance of diseases such as malaria (Kleinschmidt et al., 2000). High and low risk malaria areas were identified through use of a GIS, which combined environmental information on local rainfall and vegetation cover as well as reported malaria cases.

In the US, the State of Pennsylvania is using a GIS based surveillance system to track the West Nile virus, which is transmitted by mosquitoes. It allows storage of field and lab data generated by numerous state agencies. Feedback from the data is then made available to staff who collect the data, to key decision makers; and the public so that appropriate preventative decisions can be made (Pensylvania West Nile surveillance program, 2002).

Other features of a GIS allows for an analysis of social risk factors such as socio-economic status. These are often drawn from the census or they may be obtained from more specialized data sources. In this research project we use the Montreal City housing database.

Another specialized feature of a GIS includes a systematic examination of spatial distributions of disease (Stone, 2000, Lee et al., 2001). Stone for example showed the utility of coupling the GIS and spatial scan statistics to evaluate whether there were any significant spatial patterns among TB cases. He found specific areas in which incidence of TB cases was high. This information was then used to track spatial trends in TB cases over several years.

Most standard GIS software contain routines to calculate distances between two points, for example point A could be the hospital and point B the residential address of a patient. Bamford identified gaps in provision of remote health services to remote areas by evaluating the minimum road distances traveled by people to access health services (Bamford et al., 1999). There are also routines that can calculate area.

Query routines allow the user to explore specific features of the dataset with considerable efficiency (Vine et al., 1998). These features employ the same SQL (standard query language) familiar to users of database software like Access, Oracle or FoxPro. In figure 5 for example we have several data sources (including layer 2= case distribution and layer 3 =census tract data).

Viewing the several layers as overlays we can, at a click of an information button for the case/control or using a summary statistics wizard, view all socio-economic data for a given tract without carrying out a the long process of a questionnaire or survey of the cases to obtain demographic and socio-economic data on them. The major disadvantage of this approach is that ecological data is attributed to individuals; however this is a constraint that can be minimized by employing data sources at a finer level of analysis such as the Montreal City housing database.

In Canada every six-digit postal code has well controlled geographic co-ordinates (latitude and longitude), provided as a point by Statistics Canada. In an urbanized area, a six-digit postal code often corresponds to a city block or an apartment building, divisions much smaller than the five-digit US postal zip code. Any set of postal codes can be matched to instantly generate the geographic locations. Figure 7 shows the Island of Montreal TB cases (blue points) plotted by the six digit postal code and overlayed on a layer that shows boundaries of the 29 former municipalities as of 1999. Cases shown in red reported within the City of Montreal were plotted with greater precision by using the full street address. Several other TB studies in Africa and in the US have used a GIS to geo-code their cases and evaluate their distribution (Beyers et al., 1996, Drga et al., 2000, Stone, 2001).

H) Progress in the application of GIS to the study of TB

Advances have been made in application of GIS to urban incidence of TB, and five recent studies point to problems in appropriate scale of analyses and in choice of definitions of crowding. The Montreal study is an attempt to take one further step toward meaningful spatial analyses.

High-incidence TB neighborhoods and corridors in Virginia were identified by using a GIS approach. Specific corridors were found to be associated with specific immigrant groups (Drga et al., 2000, Dennis et al., 2001). They grouped census blocks and identified groups which were at highest risk. Following this they were able to relate TB incidence to immigrant concentration and to recognize that the pattern of high-risk block groups correlated with blocks that had a high proportion of immigrant population. Strengths of this study were the creation of visual aids (maps) to help managers justify localized programs of control as opposed to region-wide efforts (Drga et al., 2000, Dennis et al., 2001). A similar approach could be adopted in Montreal where certain areas may merit greater efforts in terms of education, awareness, and access to diagnosis and treatment.

Researchers from the US Centers for Disease Control linked TB cases reported 1987-1993 to age and race-specific socio-economic indicators at the zip code level (Cantwell et al., 1997). Variables examined included median persons per room, median household income, percentage of persons living below poverty line, percentage receiving public assistance, percent unemployed, and percentage of high school graduates. Significant elevations in adjusted TB incidence were associated with zip codes with lower rankings on any of the six indicators. Strengths of the study included adjustment for socio-economic variables to help explain the factors associated with increased risk in minority populations and the use of several alternative measures of socio-economic status as proxies for poverty.

Several points of the study could be improved, including using more refined measures of their socio-economic variables at a finer unit of analysis than the zip code, which encompasses a very large area. This would avoid assigning group estimates of covariates to all individuals in the area. Their measure of crowding based on median persons per room at the zip code could be improved on, as it is likely to vary widely from one small neighbourhood to another. The aggregation may mask an important effect and an important share of the variance. Assignment of socio-economic values to persons with TB, if it is based on median values in a zip code may underestimate the true magnitude of the association between socio-economic status and TB as TB patients may differ systematically from others within the population (Cantwell et al., 1998, Comstock et al., 1992).

Acevedo-Garcia evaluated the association between TB incidence and zip code-level attributes in New Jersey, 1985-1992 (Acevedo-Garcia, 2001). Zip codes were characterized by the distribution across the block groups of the following: poverty, crowding, proportion of dwellings rented vs. owned, racial segregation, residential mixing with immigrants, and population density.

A crowded housing unit was defined as one that contained one or more persons per room. The author documented an association between high TB incidence and poverty, overcrowding and rental housing, primarily among African Americans. They suggest that cases among African Americans and Hispanics may be due to risk factors that facilitate TB transmission. One of this study's strengths was the use of an ecological design to show that risk can be identified in ecological terms. Given the large unit of analysis (US zip code) this may also be a weakness of the study since ecological estimates are more sensitive than individual level estimates to sources of bias, such as confounding. Therefore by not adjusting or making any effort to address the issue of confounding by unknown factors, such as AIDS incidence, the estimates resulting from this study may be biased. Their definition of substandard housing was relatively crude, and may not be a very accurate measure.

Incident cases of TB in New York City in 1984 and 1992 were mapped to over 5000 census block groups (Barr et al., 2001). By using the census block group as their unit of analysis, they improved on the studies carried out by Acevedo-Garcia and Cantwell et al. where the unit of analysis was a US zip code. A zip code encompasses a much larger area than a census block group and therefore even though both designs were ecological, the inferences in the Barr study refer to much smaller areas (neighbourhoods as opposed to zip code areas).

They evaluated the association between TB incidence and poverty at the block group level, while adjusting for distribution of covariates including AIDS incidence, age, race, and foreign birth. The attempt to adjust for AIDS incidence was another improvement. The percentage of inhabitants living below the federal poverty cut-off was independently associated with TB incidence. A crowding index was calculated as the mean number of persons per dwelling, divided by the mean number of bedrooms per dwelling; they observed association with TB incidence in univariate analysis, but not in multivariate analysis.

In ecological studies the unit of analysis is the population or the community. These studies are very useful due to low cost, convenience, and an interest in ecologic effects (Rothman et al., 1998). One of the most common concerns regarding ecological studies involves the ecological fallacy, which is the bias that may occur because an association observed between aggregate level variables may not hold true for individual level data (Rothman et al., 1998).

To ensure that our study was as valid as possible a multilevel study design was also used. That is, some of our variables were truly ecological (census tract estimators), others were obtained from a smaller level of analysis (block level estimators) and finally dwelling level estimates. This was done to try to control for several possible confounders. We are still constrained by the problem of some census tract ecological data. The census tract is a somewhat relatively arbitrary administrative unit, initially delineated to be socio-economically homogeneous and compact in shape (Statistics Canada, 2001). Therefore one of the main strengths of our study, which was facilitated by use of a GIS system, was the combination of data from various levels of analysis to create a multilevel study. A GIS is required because the various databases from which we obtained our data (census data, housing data, TB data), all had different geographically defined information.

I) Why housing data?

Recent research continues to target poverty as a key factor contributing to TB incidence. Housing attributes cannot easily be separated from income and its effects on nutrition, access to care, or social deprivation, but often low-rent housing identifies populations at greater risk. Past research also suggests the value of individual level data; such as that which may arise via use of a geographically precise array of housing variables. In Montreal, newly arrived immigrants and refugees (the high TB incidence groups) likely collect in certain housing types and locations.

The City of Montreal created a high precision housing database for internal administrative use by the planning department. It is a good example of a specialized local data source that may be useful in providing additional information on risk factors that may be associated with disease occurrence. It displays the footprint of every building and detailed information on dwellings within the former City of Montreal. It also displays city blocks, the street network, and municipal and neighborhood boundaries. It can be used for identification of place names or major street names, and analysis of dwellings. It was created using records from the land registry, fire registry, urban services, public works and building inspection services, in addition to files extracted from aerial photographs. City block level measures of residential dwelling density (dwellings/m^2 in the block) and percent lot coverage (percentage of a city block that is covered by buildings in the block) can be derived from the individual dwelling attributes.

J) Rationale and significance

This approach is particularly relevant for the study of TB, because infection is transmitted by the airborne route, and is closely linked to living conditions in the immediate habitat. It provides the opportunity to exploit local data sources, which may prove to be more useful than ecological data obtained from census data at the census tract level in explaining TB occurrence.

Advantages include access and use of the high precision City of Montreal housing GIS which enables the identification and investigation of conditions or situations which may promote the spread of TB within Montreal. We aim to show that certain areas of Montreal are at high risk but in addition to this we will extend the analysis to investigate the role of socio-economic factors, specific neighborhood and housing attributes as possible factors for this occurrence.

Major improvements of this innovative approach over traditional approaches used in previous studies include its ability to obtain a much finer and more precise measure of crowding and in addition its innovative application of a traditional epidemiological study design, namely the case-cohort approach. It will be necessary to go into further detail in the methods.

Methods

To assess the usefulness of a high precision geographic housing database to improving our understanding of local urban TB epidemiology, incident cases for the period 1996-2000 were precisely mapped. All case and control dwellings were characterized in terms of dwelling attributes available from the City of Montreal housing database (individual level) and in terms of socio-economic, demographic and housing characteristics selected from census tract level data (ecologic). Variations were interpreted using logistic regression.

a) Study definitions

From a geographical perspective, we are talking about scale: the scale of the map and the population aggregates under examination. The dwelling is the basic unit of analysis; the main dependent variable is the presence or absence of a case of active tuberculosis reported from the dwelling. Explanatory factors (independent variables) that were evaluated come from three different levels of aggregation. These are individual dwelling level variables obtained from the City housing GIS, ecological neighbourhood variables at the city block level from the City housing GIS and ecological variables obtained from census data aggregated to the census tract level.

A city block, which is a much smaller unit than a census tract (figure 6), has an average size of 16,600 m^2, an average population of 128 people and an average of 60 dwellings. Note that area and population vary greatly and do not have the same degree of standardization as that of the census tract. There are 9550 city blocks in the former City of Montreal. Census Tracts are small geographic units representing urban or rural neighbourhood-like communities created in census metropolitan areas. When first delineated they were as socio-economically homogeneous and compact in shape as possible, with an average population of 4000, but their populations range from 2500 to 8000 (Statistics Canada, 1996). There are 395 census tracts in the City of Montreal. The area of a city block can be easily calculated using GIS software.

b) Study population

i) Case population

To identify case dwellings we begin with a list of all people in the City of Montreal newly reported to have active TB between January 1st 1996 through December 31st 2000 as confirmed by the Montreal public health department and entered into their notifiable diseases database (MADO). Cases that lay within the old municipality of Montreal[2] were geo-coded (precisely mapped) by virtue of their street address at the time of reporting. Cases that occurred outside of the former city of Montreal were geo-coded by virtue of the latitude and longitude point of their corresponding six digit postal code at the time of reporting (Figure 7, in blue), however only cases that lay within the former city of Montreal (Figure 7, in red) were used for further analysis.

Of the original 902 cases, 39 were excluded as they lived in institutions (health care, correctional or religious) leaving us with 863 cases. 595 cases in the former city boundaries were precisely mapped by virtue of the street address, and 244 cases in the new districts outside of the city were mapped according to the latitude and longitude of their six digit postal code. Twenty-four cases that lay within the City of Montreal could not be located by their street address, and were therefore postal geo-coded.

The most plausible explanation for the inability to geo-code the 24 addresses in the City was that the dwellings were no longer standing in 2000. The validity of these addresses was verified from paper records at the public health department. These cases were excluded from analysis that required precise housing data, but were retained for mapping distributions on the Island (figure 7) and can be used in analysis limited to census tract level data.

ii) Control dwelling population

A suitable set of controls must originate from the same reference population as our cases. City case dwellings were obtained from a list of residential dwellings located within the City of Montreal. To ensure that control dwellings were a representative sample of the distribution of potential exposures from which the cases arose, they were randomly selected from the same list of all residential dwellings (approximately half a million) in the City housing database. Each dwelling was enumerated and a sample of 5950 control dwellings were randomly selected (Appendix 4).

This study was approached using a case-cohort design. The cohort was defined as the total number of all residential dwellings within the City of Montreal. Due to the large size of our cohort, it was impractical to undertake covariate processing for half a million records. Because of the relatively low rate of disease occurrence, much of the covariate information on disease-free subjects (our control dwellings) would be largely redundant (Prentice, 1986). Therefore, the decision was taken to obtain a sample of this control population (our sub-cohort) in the ratio of 1:10 case dwellings to control dwellings. 5950 control dwellings were selected and mapped.

The final dataset for analysis therefore consisted of 6545 records, of which there were 595 case dwellings and 5950 control dwellings.

c) Variables used

Bearing in mind the level of aggregation which could be obtained, let us look at the variables with special attention to the measures of crowding, and to the variables which can be introduced as characteristics of individual dwellings (cases or controls) without aggregation.

The variables included in this study (table 1) are of three types: (1) variables which characterize the individual dwelling unit, all drawn from the City housing GIS; (2) characteristics of the city block, compiled from both census and municipal sources (figure 6); and (3) average population characteristics from census tract. Several variables from the last two sets are derived from the census, some are aggregated from the municipal data for individual buildings and dwellings, and some are calculated by combining variables.

Table 1: Variables, sources and levels of aggregation

Source	Level of aggregation	Variable	Units	Formulation of variables
	Census tract	Population density	people/Km^2	Census tract population / area of the census tract
	Census tract	Percent of immigrants	%	Obtained from a 20 % sample
	Census tract	Percent of recent immigrants	%	Obtained from a 20 % sample
	Census tract	Median income	$	Median income of people aged 15 years and older. From a 20% sample
	Census tract	Mean number of rooms per dwelling	rooms	Obtained from a 20 % sample
	Census tract	Mean household size	people	Obtained from a 20 % sample
Census of 1996, 20% sample	Census tract	CT crowding index(people/room)	people /room	Obtained by dividing the mean household size by the mean number of rooms
	City block	Floor area per person	m^2 / person	Number of people in a city block / total floor area in a city block. An estimate of the block population was obtained by summing the population estimates of all six-digit postal codes that lie within a city block. The floor area of the block was obtained by summing (dwelling footprint area x number of stories) (figure 8)
City housing GIS and census	City block	Block population density	people/Km^2	Number of people in a city block / city block area. An estimate of the block population was obtained by summing up the population estimates of all six-digit postal codes that lie within a city block
	City block	Block residential dwelling density	dwelling / Km^2	Number of dwellings / area of a city block.
	City block	City bock lot coverage	%	Sum of the areas of building footprints in a block, divided by the area of the block
	Dwelling	Dwelling age in the year 2000	years	Age of building assigned to all dwellings within
	Dwelling	Dwelling value	$	Value of the building / number of dwellings in the building
City housing GIS	Dwelling	Stories		The number of stories of the building in which dwelling lies

The primary use of variables from the City housing GIS was to obtain individual level data on variables that may be important predictors of TB occurrence. By combining and manipulating the City housing database with data from the Census Canada, an estimate of the population density at the city block level and the floor area per person will be obtained. This will give us a measure of both residential density and of crowding at a very fine unit of analysis (the city block), thus enabling us to evaluate them as risk factors. Socio-economic values from the Census will allow us to evaluate poverty and the effect of recent immigrants populations.

d) Data sources

Data for this study was obtained from the following three sources 1) The MADO, 2) The City housing GIS, and 3) Census Canada. The MADO is the only data source that was created specifically for disease surveillance with systematic collection and reporting of data. The City housing GIS is updated yearly and was designed for use by the municipal planning department, while Census Canada data is created every five years. The variables of interest to us are provided from a 20 percent sample. The block population is updated yearly and based on a 100% sample. Advantages and limitations with the use of these datasets are evaluated in the discussion.

The City of Montreal electronic planning housing geo-database[1] covers only the former municipality of Montreal as bounded in 1999. This amounts to approximately 57 % of the present population of Montreal and 72% of the TB burden on the island of Montreal for the period 1996 – 2000. The housing stock is biased to inner-city distributions.

It is a high precision database that provides detailed information on dwelling attributes in the City of Montreal. Its layers of electronic maps known as themes allow us to view the various housing and other locality attributes. The theme is accompanied by a table of data and those of interest to us contained dwelling characteristics such as number of stories, number of dwellings in each building, dwelling age, value of building, area of the building and city block (for calculation of residential dwelling density per city block). The database consisted of 468,130 residential dwellings and the municipal data were updated as of January 2000.

The "address-view theme", obtained from the City housing GIS was modified by Guido Guerra to search more efficiently for French language street names. It was used to geo-code MADO case addresses that lay within the old municipality of Montreal. Geo-coding was carried out using ArcView 3.2 GIS software, generating a layer of points, each with geographic co-ordinates. They can be viewed only in relation to a map projected in the same way (Moore et al., 1999), with the same degree of precision and detail as the City housing GIS. Coordinates are in the Modified Transverse Mercator, NAD83 (Zone 8) projection, but can be converted to latitude-longitude or any other projection. Following geo-coding, the actual address numbers and street names were removed from the table.

The postal code conversion file from Census Canada contains latitude and longitude coordinates for points which represent the approximate location of each Canadian six-digit postal code (Statistics Canada, 2000). These coordinates were projected from latitude/longitude to match the Modified Transverse Mercator, NAD83 (Zone 8) map. The points were used to represent cases that occurred outside the old municipality of Montreal. In most cases they provide a central point for the block in which the dwelling is situated, and are more accurate for a large building (which would have its unique postal code and least accurate in low-density suburbs and fringe areas.

The 1996 Census obtained from Statistics Canada (Statistics Canada, 1996) under the Data Liberation Initiative to university libraries, was the source for socio-economic and demographic variables, aggregated to the census tract level. The census was the sole source readily available for estimating distributions of the recent immigrant population and their residential concentrations by country of birth. Note that several variables come from a 20 percent sample (table 1).

In order to align census data with the City housing GIS the geographic coordinates of the digital cartographic file of the Montreal area census tracts were converted from latitude/longitude to Modified Transverse Mercator, NAD83 (Zone 8) projection.

e) Data management and analysis

In preparing the data, I had to deal with several problems that naturally arise when we combine several sources of data. (1) I had to adjust crude census tract boundaries created by Statistics Canada to the more precise municipal SIURS, so that the boundary lines did not slice through buildings. (2) I had to verify the heights of several buildings which were reported as single management units, notably in a public housing project which mixed high and low rise buildings, and a large coop housing project with dozens of buildings. (3) The distribution of each variable was examined to identify possible trends and to look for possible outliers and missing data. The only serious problem of missing data was the variable dwelling age which had missing age information for 8% of dwellings. To deal with this, the variable was categorized with an additional category for missing information. This enabled me to maintain a full dataset for analysis as well as examine how categories of dwelling age related to TB occurrence. (4) A correlation analysis was performed between all pairs of independent variables and highly inter-correlated variables (>0.70) were dropped. (5) The linearity of all variables was examined and several variables were categorized (appendix 5, table 5) as it was not appropriate to run them as continuous variables i.e. they were not linear in the logit (appendix 1-3).

There were four main data analysis steps. First, census and housing data were attributed to case and control dwellings using ArcView 3.2 GIS software. This software was also used to explore for outliers in the data and to create some calculated variables. Second, the data was imported into SAS version 8.02 for descriptive statistics. Third, the usual sequence of univariate and multivariate logistic regression was used to model the relationship of TB occurrence to dwelling-level attributes from the Montreal Housing database. Both univariate and multivariate effect estimates were generated. Univariate analysis was first used to model TB occurrence to each of the independent variables, with the dependent variable being the presence or absence of a TB case within a given dwelling. Fourth, several census variables (imputed to cases) were forced into the model, and estimates were re-examined. These included percent recent immigrants, median income, people per room, and floor area per person, as previous studies have identified them as important predictors.

An appropriate multivariate model was then developed while examining possible confounding and effect modification by fitting relevant housing variables alone, using backward stepwise regression beginning with variables significant at $P>0.20$ from the univariate analysis. This multivariate model was assessed using deviance testing and the Hosmer-Lemeshow goodness of fit statistic (Hosmer et al., 2000).

Results

A. Characteristics of case and control dwellings

By straightforward comparison of case and control dwellings, one variable at a time, we discover some initial contrasts. Some of the differences are small but they raise new questions and help design the multivariate model. As anticipated, relative to control dwellings, case dwellings are disproportionately located in census tracts of low income, in census tracts with higher percentages of recent immigrants, in tracts and in city blocks with higher population densities (table 2). This is true regardless of whether residential density is measured in terms of persons or dwellings per square km. Case dwellings are slightly more crowded, as measured by the census tract index (number of people/room) and by the finer-grained estimate for city blocks (floor area per person).

Relative to control dwellings, case dwellings are disproportionately located in census tracts that have a lower median income of persons 15 years and older. Where we use individual dwelling values obtained from the municipal housing data, case dwellings have a lower median value than control dwellings (table 2). The dwelling value is a precise measure of each dwellings actual valuation, as obtained from the municipality of Montreal. The median incomes for case and control census tracts are not as divergent as one might expect. The reason for this is that the area studied, the old municipality of Montreal, excludes most of the wealthiest census tracts on the Island and therefore the median incomes for census tracts with control dwellings are representative of the former city of Montreal.

We need to bear in mind the rather low variance which results from two factors: the rather standard and unique terrace housing stock of Montreal and the fact that the old municipality excluded most of the wealthiest census tracts. In half of all census tracts in the city, mean dwelling size ranges between four and five rooms, household size between 2.1 and 2.3 persons. Half the buildings date from the era 1935-1970, half are one or two stories tall, and three-quarters of dwellings are valued over $20,000. Half of the city blocks have less than 37% lot coverage.

If we turn to the data for individual dwellings (from the municipal housing database), relative to the set of controls, dwellings with cases of TB are associated with higher population densities for both the census tract and city block (table 2). If density is measured in the form of the number of dwellings the same is true.

Contrary to our expectations case dwellings were more recently constructed than control dwellings. The great majority of residential dwellings in the old municipality are two-story or three-story, often flats occupying entire floors, known as duplexes or triplexes. Only approximately ten percent of the control dwellings were located in buildings taller than three stories.

There are 395 census tracts in the City of Montreal. The 595 case dwellings are distributed among 215 of these census tracts, while 5950 control dwellings lie within 318 of these different census tracts. There are 9550 city blocks in the former City of Montreal. 595 case dwellings lie within 477, while 5950 control dwellings lie within 3385 of these city blocks.

Case dwellings are located in census tracts that have a higher percentage of immigrants as we would expect. This is consistent with the characteristics of the case population (table 3) where we see that the percentage of immigrant cases is 83%. The percent of recently arrived immigrants (since 1990) is also much higher among case dwellings than control dwellings.

Table 2: Comparison of case and control dwellings

Level of analysis	Variable	Units	Case dwellings (595) Median (IQR)	Control dwellings (5950) Median (IQR)
Census tract	Population density	people/Km^2	11187 (6725, 16227)	9406 (6111, 14228)
	Percent recent immigrants	%	10.18 (5.54, 19.90)	5.81 (2.8, 10.2)
	Percent immigrants	%	38 (20, 53)	22 (14, 38)
	Median income of persons 15 years and older	$	13120 (11276, 15221)	14953 (12684, 17482)
	Average number of rooms per dwelling	rooms	4.5 (4.1, 4.9)	4.7 (4.4, 5)
	CT crowding index	people /room	0.48 (0.43, 0.57)	0.44 (0.41, 0.48)
	Household size	people	2.2 (2.0, 2.5)	2.1 (1.9, 2.3)
Dwelling	Dwelling age in the year 2000	years	42 (34, 51)	46 (37, 72)
	Dwelling value	$	24573 (16125, 45200)	30766 (20210, 49700)
	Block residential dwelling density	dwelling / Km^2	10653 (6365, 15510)	9094 (5340, 13743)
	Block population density	people/Km^2	21155 (12561, 31796)	17231 (10003, 25986)
	Stories		3 (2, 3)	2 (2, 3)
	Floor area per person	m^2/ person	42 (32, 58)	46 (36, 61)
City block	Lot coverage	%	38 (29, 46)	37 (28, 45)

The 24 case dwellings that seem to have been demolished were excluded from the case control analysis. With the aid of their postal codes, these 24 dwellings were mapped and assigned to their respective census tract and data imputed to them. A comparison was made to evaluate the characteristics of these dwellings (table 2b). They had similar characteristics to the other case dwellings as explained earlier.

Table 2b: Comparison of all case and control dwellings

		24 case dwellings	**Case dwellings (595)**	**Control dwellings (5950)**
Variable	**Units**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**
Population density	people/ Km^2	14381 (8443, 16868)	11187 (6725, 16227)	9406 (6111, 14228)
Percent recent immigrants	%	9.13 (2.73, 14.49)	10.18 (5.54, 19.90)	5.81 (2.79, 10.18)
Percent immigrants	%	36 (16, 43)	38 (20, 53)	22 (14, 38)
Median income of persons 15 years and older	$	12784 (11258, 15948)	13120 (11276, 15221)	14953 (12684, 17482)
Average number of rooms per dwelling	rooms	4.5 (4.1, 4.9)	4.5 (4.1, 4.9)	4.7 (4.4, 5)
CT crowding index	people /room	0.48 (0.43, 0.54)	0.48 (0.43, 0.57)	0.44 (0.41, 0.48)
Household size	people	2.2 (1.9, 2.5)	2.2 (2.0, 2.5)	2.1 (1.9, 2.3)

B. Characteristics of the actual case population

The TB case load in Montreal consists of a high percentage (83%) of immigrants (table 3a). Recent immigrants in Canada (less than 5 years), make up 25% while those less than 10 years make up nearly half of the TB case load (table 3b).

TB case occurrence is generally somewhat more frequent in men than in women, more so among the immigrants.

Table 3a: Case description by country of birth and gender

			Male		Female	
Country of birth	**Number**	**Percentage**	n	%	n	%
Total cases	595	100	329	55	266	45
Canadian-born	97	16	51	53	46	47
Foreign-born (immigrants)	493	83	276	56	217	44
Unknown origin	5	1	2	40	3	60

Table 3b: Immigrant case description by year of arrival

Interval between arrival and diagnosis (years)	**n (493)**	**% of all cases**
0-5	152	25.5
6-10	127	21.3
11-20	111	18.7
>21	71	11.9
Unknown	32	5.4
Total	493	82.9

C. Correlations between variables

To develop a multivariate model, we first need to test for inter-correlations, examine the suitability of variables being used as continuous and to consider appropriate categorization for those that are not linear in the logit. We also need to examine any remaining risk of confounding.

Several variables are to be dropped in order to avoid multi-collinearity. These include census tract population density, and block population density. Block population density is strongly correlated with block residential density (Table 4). Since we are still evaluating the housing variables, block population density is initially kept to evaluate the validity of the estimates in univariate analysis (i.e. Is the trend following univariate analysis in the same direction as what we would expect), but will be dropped following this.

Average number of rooms per dwelling strongly correlated with household size. High lot coverage is strongly associated with residential density measures (0.39 at block level and 0.59 at census tract level). It is encouraging to confirm the correlations among lot coverage and the measures of density at the two levels of observation (city block level and census tract (0.42, and 0.45)). This suggests that the two data sources are providing reasonably consistent and perhaps substitutable estimates for measures of density.

Despite inter-correlations between percent of recent immigrants and people per room these two variables were left in as they are important known predictors and we will have to take care and examine confounding effects.

Table 4: Correlation matrix

	Pop den.	R.im migra nts	Med inc.	Rooms /dwell	Househ old size	People / room	Dwell age	Dwell value	Res. density	Block pop. Den.	Story	Floor area/ person	Lot covera ge
Census tract population density	1.00	0.37	- 0.37	-0.48	-0.17	0.29	0.19	-0.18	0.42	0.45	0.19	0.00	0.39
% Recent immigrants		1.00	- 0.44	-0.52	0.17	0.76	-0.08	-0.17	0.27	0.30	0.17	-0.03	0.13
Median income			1.00	0.48	0.00	-0.47	-0.19	0.39	-0.28	-0.32	-0.06	0.03	-0.36
Average rooms per dwelling				1.00	0.61	-0.29	-0.18	0.30	-0.47	-0.38	-0.27	-0.03	-0.38
Average household size					1.00	0.58	-0.35	0.13	-0.29	-0.13	-0.18	-0.09	-0.35
People per room						1.00	-0.23	-0.14	0.13	0.24	0.05	-0.07	-0.03
Dwelling age							1.00	-0.13	0.05	0.09	-0.08	0.05	0.39
Dwelling value								1.00	-0.27	-0.22	-0.04	0.06	-0.23
Block residential dwelling density									1.00	0.73	0.36	-0.03	0.59
Block population density										1.00	0.26	-0.17	0.56
Stories											1.00	0.07	0.11
Floor area per person												1.00	0.05
Lot coverage													1.00

D. Categorization of variables

Multivariate logistic analysis can handle both linear and categorical variables, if appropriately formulated. Suitable for use as continuous variables, verified through linearity checks (Appendix 1-3), are percentage of recent immigrants, median income, and census tract crowding (people/room).

Appropriate categories for all other variables were determined by initially categorizing all variables using the quartiles obtained by running distributions on the control set. Logit plots were then done to verify their suitability as being run as continuous variables, as well as to determine appropriate categorizations cut offs (Appendix 1,2, 3, 5). Based on their crude univariate odds ratio pattern in relation to their reference category a decision was then made to re-categorize variables as follows (table 5: see also appendix 5).

The distribution of buildings (and therefore of dwellings) by their age is affected by a very strong building cycle, (Figure 9) common to all of North America. This cycle was driven in part by ebbs and flows in the economy, and the need for post war housing. The variables dwelling age and story were therefore initially categorized by using natural breaks in their distribution (Figure 9).

Table 5: Final categorizations of variables

	Cat.	Description of category	% in cat.	~ Quartile location
Census tract rooms per dwelling	1	>5	28	>75
	2	> 4 and <5	60	25-75
	3	1-4	12	<25
Census tract household size	1	>2.3	23	>75
	2	>2.1 and <=2.3	17	50-75
	3	<=2.1	60	<50
Dwelling age	1	missing age*	8	missing
	2	>65	26	>75
	3	>30 and <=65	52	25-75
	4	1-30	14	<25
Dwelling value	1	>20000	74	>25
	2	<=20000	26	<25
Block residential dwelling density	1	>14000	25	>75
	2	>9300 and < =14000	25	50-75
	3	<=9300	50	<50
Block population density	1	missing*	5	missing
	2	>27000	23	>75
	3	>175000 and <=27000	24	50-75
	4	<=175000	48	<50
Block floor area per person	1	missing*	5	missing
	2	>61	24	>50
	3	>35 and <61	49	25-50
	4	<=35	22	<25
Stories	1	>3	11	
	2	3	38	
	3	1-2	51	
Lot coverage	1	>37%	50	>50
	2	<=37%	50	<50

* Missing category for dwelling age exists due to several dwellings lacking a dwelling age. Missing category for block dwelling density and block floor area exists due to a lack of population estimates in several blocks.

E. Crude odds ratio interpretation

From an initial interpretation of crude odds ratios, all have the relationship (direction) we would anticipate, and almost all are statistically significant as they are within the 95% confidence interval (Table 6 and figure 10).

There is a 45% increase in risk in a census tract that contains a 10% higher percentage of immigrants. This risk is even higher for recent immigrants (the risk is doubled). As median income of the census tract increases by 1000$, there is a decrease in risk by 13%. Effects from the actual dwelling value indicate that there is a threshold value at which dwellings are more likely to be case dwellings, approximately 20,000$ (dwellings within the lowest quartile). Above this threshold, risk decreases by half. All of the crowding and density variables used indicate that dwellings that are more densely occupied or dwellings located in areas that are more densely populated are at a greater risk of housing a case of TB. Relative to older dwellings (65 years and older), younger dwellings are at an increased risk. Dwellings located in buildings with >3 stories are at a higher risk of being a case dwelling.

These effects reinforce the hypotheses of poverty, density, crowding and immigration, that is the 'urban' nature of the distribution, already recognized in the conception of the study. Somewhat more interesting, the introduction of sources on individual dwellings gives consistent results (value, density) but adds three new observations: (a) the apparent threshold value of dwellings, (b) the higher incidence of TB in newer housing, and (c) higher incidence in taller buildings. Since control dwellings were randomly selected from all dwellings in the City (Appendix 4) differences between case and control dwellings are not the result of underlying dwelling concentrations in specific areas or buildings.

Table 6: Crude (unadjusted) odds ratios for tuberculosis

Variable	Category	Interpretation of OR	Crude OR	95 % CI	
CT percent foreign born	Continuous	Per 10% increase	1.45	1.38	1.52
CT percent Recent F. born	Continuous	Per 10% increase	2.21	2	2.44
CT median income	Continuous	Per 1000$ increase	0.87	0.85	0.89
CT crowding index (people/room)	Continuous	Per 0.1 person increase	2.29	2.07	2.54
CT rooms per dwelling	1-4	In relation to the reference category	2.19	1.67	2.88
	>4 and <5		1.4	1.13	1.73
	>5 (reference)		1		
CT household size	<=2.1 (reference)	In relation to the reference category			
	>2.1 to <=2.3		1.48	1.18	1.86
	>2.3		2.08	1.72	2.53
Dwelling age	1-30	In relation to the reference category	2.24	1.65	3.05
	>30 and <=65		2.53	1.98	3.3
	>65 (reference)		1		
	missing age		1.16	0.76	1.77
Dwelling value	<=20000	In relation to the reference category	2	1.72	2.43
	>20000 (reference)		1		
Block dwelling density	<=9300 (reference)	In relation to the reference category	1		
	>9300 to <=14000		1.36	1.1	1.69
	>14000		2.14	1.76	2.6
Block population density	<=175000 (reference)	In relation to the reference category	1		
	>175000 and <=27000		1.24	1	1.55
	>27000		2.01	1.65	2.46
	missing		1.35	0.9	2.01
Block Floor space (m2/person)	<=35	In relation to the reference category	1.63	1.29	2.05
	>35 and <=61		0.91	0.73	1.13
	>61(reference)		1		
	missing		1.06	0.7	1.61
Story	1-2 (reference)	In relation to the reference category	1		
	3		1.79	1.48	2.15
	>3		2.22	1.72	2.86
Lot coverage	<=37% (reference)	In relation to the reference category	1		
	>37%		1.25	1.05	1.48

F. Examination of possible confounders

As several of the key variables are inter-correlated, it is important to consider potential confounding. In particular, with respect to three main exposure variables (recent immigrants, median income and floor area per person). To do this a logistic regression was carried out with one main exposure variable and the effect of adding an additional variable (as measured by the change in the ß coefficient) was examined to determine whether the additional variable acted as a confounder. Variables that caused a change in the β co-efficient by >15% were classified as confounders.

First, median income confounded the relationship between recent immigrants and the outcome (Table 7a). The relationship between recent immigrants and the outcome variable was also confounded by crowding at the census tract level. This was expected as recent immigrants and people per room were strongly correlated in the correlation analysis (table 4).

Table 7a: Confounding of recent immigrants and case dwelling status

Variable evaluated as a possible confounder	Crude β	New foreign born β estimates when adjusted for variable
Percent recent immigrants per 10% increase	**0.37**	
Median income per 1000$		0.30
Dwelling age in the year 2000		0.34
Block Population density		0.35
Floor area per person		0.35
Household size		0.38
Block residential dwelling density (using dwelling estimates from SIURS)		0.34
% Lot coverage per city block		0.37
Average number of rooms per dwelling		0.36
Story		0.35
CT average value of dwelling		0.42
Siurs dwelling value		0.34
CT crowd (people/room)		0.17

Recent immigrants confounded the relationship between median income and the risk of being a case dwelling (Table 7b). The relationship between median income and the outcome variable was also confounded by crowding at the census tract level.

Table 7b: Confounding of median income and case dwelling status

Variable evaluated as a possible confounder	Crude β	New median income β estimates when adjusted for variable
Median income per 1000$	**-0.14**	
Percent recent immigrants per 10% increase		-0.07
Dwelling age in the year 2000		-0.15
Block population density		-0.13
Floor area per person		-0.13
Household size		-0.13
Block residential dwelling density (using dwelling estimates from SIURS)		-0.12
% Lot coverage per city block		-0.14
Average number of rooms per dwelling		-0.14
Story		-0.13
Siurs dwelling value		-0.12
Ct Crowd (people/room)		-0.05

*Highlighted are potential confounders that will need to be adjusted for

The variables recent immigrants, household size and crowding at the census tract level all confounded the relationship between floor area per person and the risk of being a case dwelling (Table 7c).

Table 7c: Confounding of the effect of floor area and case dwelling status

Variable evaluated as a possible confounder	Category	Crude ß[4]	New floor area per person ß estimates when adjusted for variable
Floor area per person	<=35	**-0,59**	
	>35 <=61	**-0,49**	
	missing	**-0,43**	
Percent recent immigrants per 10% increase	<=35		-0,31
	>35 <=61		-0,19
	missing		-0,12
Median income per 1000$	<=35		-0,52
	>35 <=61		-0,34
	missing		-0,39
Dwelling age in the year 2000	<=35		-0,49
	>35 <=61		-0,31
	missing		-0,35
Block population density	<=35		-0,44
	>35 <=61		-0,20
	missing		-12,00
Average household size	<=35		-0,40
	>35 <=61		-0,28
	missing		-0,35
Block residential dwelling density	<=35		-0,56
	>35 <=61		-0,40
	missing		-0,48
% lot coverage	<=35		-0,60
	>35 <=61		-0,50
	missing		-0,48
Average # of rooms / dwelling	<=35		-0,63
	>35 <=61		-0,58
	missing		-0,38
Story	<=35		-0,61
	>35 <=61		-0,64
	missing		-0,37
CT average value of dwelling	<=35		-0,58
	>35 <=61		-0,49
	missing		-0,36
Siurs dwelling value	<=35		-0,56
	>35 <=61		-0,46
	missing		-0,35
CT crowd (people/room)	<=35		-0,11
	>35 <=61		0,14
	missing		-0,07

[4] β co-efficients are in relation to the reference category of >61 m^2 /person.

G. Sequential automated model selection

Following the analysis of confounding variables that are forced in the model include recent immigrants, median income, crowding at the census tract level and the city block level (floor area per person).

An appropriate model was selected for further analysis by performing a backward logistic regression selection procedure, with a stay criterion of 0.20. The resulting model is shown below (table 8).

Table 8: Model selection via backward logistic regression

Variable	Category	Backward selection	95 % CI for adjusted model	
Recent immigrants per 10% increase	Continuous	1.42	1.15	1.74
Med. inc. per 1000$ increase	Continuous	0.96	0.93	0.99
Crowding per 0.1 person increase	Continuous	1.39	1.12	1.73
Rooms	1-4	0.74	0.52	1.04
	>4 and <5	0.98	0.77	1.26
	>5 (reference)	1.00		
Dwelling age	1-30	2.43	1.72	3.44
	>30 and <=65	1.79	1.36	2.35
	>65 (reference)	1		
	missing age	1.13	0.72	1.77
	<=20000	1.27	1.03	1.57
Dwelling value	>20000 (reference)	1		
Floor area per person	<=35	0.95	0.72	1.25
	>35 and <=61	0.81	0.64	1.03
	>61(reference)			
	missing	0.92	0.59	1.43
Story	1-2 (reference)	1		
	3	1.26	1	1.57
	>3	1.55	1.15	2.1
	<=37% (ref.)	1		
Lot coverage	>37%	1.25	1.03	1.52
Aikaike information criteria		3694.08		
-2 log likelihood (deviance)		3662.02		

H. Model assessment

The model selected for further analysis (table 8) was examined for possible interactions, goodness of fit and its overall suitability via deviance testing (table 9). It was also compared to a model with census data only (model 4 table 9) and with housing data only to appraise the value added by the individual housing data.

None of the interaction terms tested (models 2 and 3 table 9) were found to be significant predictors and they did not contribute much to the model, as indicated by the minor change in the AIC (model 1, table 9). This implies that there was no effect modification between the interaction terms on the case control status.

When the -2 Log Likelihood for the following model pairs were compared (Model 1 and model 2) and (model 1 and model 4) all were found to have a chi square result that was not significant. In addition the AIC values barely changed for models that included an interaction term. In fact the AIC values were marginally higher in models that included an interaction term. This too, indicates that the interaction term did not improve the model and the fit was slightly worse. We can therefore infer that 1) the effect of crowding on being a case dwelling is the same regardless of the percentage of recently arrived immigrants in the census tract and 2) the effect of crowding on being a case dwelling is the same in census tracts of different income levels.

Value added by City housing variables

Model 1 (table 9) which included housing variables fits much better than model 4 (table 9) with census variables only as it has a much lower AIC value. It was also better than model 5 with housing data and recent immigrants only. Model 5 however was still a better model than that with census data only (model 4). This suggests that both census and individual housing level data were useful in explaining TB occurrence and that the variable recent immigrants is a very strong predictor.

A significant chi square result was obtained after comparing the -2 Log Likelihood for models 1 and 4. In other words, the housing variables improved the model.

Goodness of fit

Model 1 (table 9) was assessed using the Hosmer-Lemeshow Goodness of fit, which tests the effectiveness of our model in describing the outcome variable. The test value had a Chi square of 4.55 and a p value of 0.80. The null hypothesis for this test is that the fit is good. Since we can not reject the null hypothesis (p value is greater than 0.05), we conclude that the fit is reasonable.

Table 9: Model assessment * Highlighted=Interaction terms tested

Variable	Model with no interaction	Interaction 1	Interaction 2	Census only	City only
	Model 1	Model 2	Model 3	Model 4	Model 5
Median income per 1000$ increase	x	x	x	x	
Recent immigrants per 10% increase	x	x	x	x	x
Rooms /dwelling (Per 0.1 person increase)	x	x	x	x	
Recent immigrants * (rooms/dwelling)		Non significant			
Median income * (rooms/dwelling)			Non significant		
Ct rooms	x	x	x	x	
Block floor area	x	x	x		x
Dwelling age	x	x	x		x
Dwelling value	x	x	x		x
Story	x	x	x		x
Lot coverage	x	x	x		x
Difference in aikaike information criteria (AIC) from model 1		0.04	1.75	46.96	18
Difference in -2 log likelihood (deviance) from model 1		-1.96	-0.25	66.88	26

Comparison of the crude and adjusted odds ratios (table 10, Figure 11) indicates that indeed a multivariate model was required in order to properly explain the effects of the different variables on the outcome. When adjusted for all other variables in the model the effects of recent immigrants are strongly and independently associated with the occurrence of TB. It explained a larger amount of the variability in TB occurrence than the other measures. This is not surprising and indeed was anticipated a priori. The question we are trying to address is the relevance of housing conditions to occurrence.

Crowding at the census tract level was an important predictor; however at the block level crowding was important only in univariate analysis but was not a significant factor once adjusted for all other variables in the model. From the city block lot coverage which is a proxy for the dwelling density and indirectly the population density, residential density does appear to be a significant predictor. It provides some indication of the level of gathering of people and perhaps exposure in an urban environment outside the family.

Other significant predictors included median income, dwelling age, dwelling value and the number of stories of the building in which the dwelling was housed. Dwelling value may be a substitutable measure of socio-economic status and its advantage is that it is a finer individual level measure.

Table 10: Comparison of the crude and adjusted odds ratios

Variable	Category	Crude OR	Adjusted OR (final model)	95 % CI for adjusted model	
Recent immigrants per 10% increase	Continuous	2,21	1,42	1,15	1,74
Median income per 1000$ increase	Continuous	0,87	0,96	0,93	0,99
Crowding per 0.1 person increase /room	Continuous	2,29	1,39	1,12	1,73
Rooms	1-4	2,19	0,74	0,52	1,04
	>4 and <5	1,4	0,98	0,77	1,26
	>5 (reference)	1	1,00		
Dwelling age	1-30	2,24	2,43	1,72	3,44
	>30 and <=65	2,53	1,79	1,36	2,35
	>65 (reference)	1	1		
	missing age	1,16	1,13	0,72	1,77
	<=20000	2	1,27	1,03	1,57
Dwelling value	>20000 (reference)	1	1		
Floor area per person	<=35	1,63	0,95	0,72	1,25
	>35 and <=61	0,91	0,81	0,64	1,03
	>61(reference)				
	missing	1,06	0,92	0,59	1,43
Story	1-2 (reference)	1	1		
	3	1,79	1,26	1	1,57
	>3	2,22	1,55	1,15	2,1
	<=37% (ref.)	1	1		
Lot coverage	>37%	1,25	1,25	1,03	1,52

Discussion

The use of a high precision dwelling referenced GIS enabled us to precisely map all incident TB cases in the former city of Montreal and to create a crowding index at the city block level. Disease occurrence in relation to dwelling attributes was characterized and associations between TB and variables that distinguish poverty, crowding and birthplace were modeled.

Socio-demographic predictors at the census tract level included crowding, median income, and percentage of immigrants. Several dwelling level risk factors were also independently associated with TB occurrence. These included dwelling age, dwelling value and the number of stories in the building. These are all indirect proxies for the socio-economic status and thus reinforce the association of low socio-economic status and occurrence of TB. However local crowding (floor area per person) was not found to be a significant risk factor.

The risk of becoming infected with TB is generally associated with factors related to the environment such as the level of exposure while the risk of developing active TB is often related to host factors such as malnutrition, weak immune system. It should be noted that no extremes of crowding were found in this study. The manner in which population density can be thought to play a rôle in TB transmission could be due to increased encounters in the urban habitat. Crowding can also play a role in terms of ventilation and transmission within a household or dwelling complex where people share the same airspace. Sharing breathing space for long periods of time in dwellings with poor and inadequate ventilation would cause the risk to be elevated.

The variables used to indicate crowding included people per room at the census tract level, and floor area per person at the city block level. Previous studies carried out in the US indicated that crowding was associated with an increased risk of TB. However, these studies (Cantwell et al., 1998, Acevedo-Garcia, 2001) were both ecological in nature, with data being imputed from the zip code level, a much larger area than a census tract. In a similar study (Barr et al., 2001) carried out at a much finer level of analysis (the census block group) crowding was associated with TB incidence in univariate analysis but was not statistically significant when adjusted for other variables. This was similar to what we observed at the city block level.

Crowding as measured by people per room at the census tract level was independently associated with TB occurrence in both univariate and multivariate analysis. When adjusted for other variables in the model, the magnitude of the effect decreased but was still statistically significant and as the crowding index increased (i.e. a 0.1 person/room increase) there was an increased risk.

However, while localised room crowding as measured by the variable floor area per person, and averaged at the city block level, was a significant predictor in univariate analysis, when adjusted for other potential predictors this association became non significant. In univariate analysis there appeared to be a threshold whereby dwellings which had floor area below this threshold were at an increased risk. Larger dwellings with more floor area per person were at a lower risk. This was similar to what Cantwell had observed in multivariate analysis using an index of crowding from the zip code level. From the confounding analysis (table 7c) we saw that the relationship was confounded by the variables recent immigrants, crowding and household size (as imputed from census data for the tract). By including these variables in the final model we have attempted to control for confounding but this results in no significant effect of localised crowding on the risk of TB. These effects of localised crowding are in opposition to what we had hypothesized. The results however are similar to the results obtained by Barr at the census block group level.

The reason why crowding as estimated from census tract level measures was important while at the city block level it was not significant may be an example of an ecological fallacy caused by an uneven distribution of crowding within census tracts. The floor area per person estimates clearly indicate that crowding within city blocks is not evenly distributed (figure 12). Therefore assigning an averaged value of crowding to all case and control dwellings may cause misleading results, especially if case dwellings tend to lie in areas of the tract that have a higher crowding index than the average value for that tract. It may also be due to the accuracy of our measure of floor area per person, which was based on calculations (figure 8) from several estimates (population from the postal code, area of block, area of dwelling).

Residential density at the city block level, as measured by proxy variable percent lot coverage, was a significant predictor of TB occurrence in both univariate and multivariate analysis. Blocks that have a higher percentage of their surface area covered by dwellings were more likely to contain case dwellings. It is the simplest and probably most reliable indicator of population density of those we tested. This supports the argument for contagion via encounter. The notion of transmission via exposure can be visualized to occur in urban environments where people tend to congregate socially, such as shopping malls, transit routes (bus/subway), and elevators to name a few.

Residential density is closely related to crowding although not conceptually identical. As anticipated, it was closely associated with TB occurrence at the block level and was a more direct and likely more reliable measure than floor area per person because it did not require as many calculations and therefore was not as prone to measurement error.

Newer dwellings (age range 1-30 and 30-65) were at higher risk than older dwellings (age range >65), even when adjusted for all other potential predictors in the model. This was in contrast with the hypothesis which assumed that older dwellings would be at an increased risk of being case dwellings. A possible explanation for this may be that newer dwellings are the housing choice of immigrants. Although we adjusted for the percentage of immigrants in the tract, since it was an ecological variable we may only obtain a partial adjustment. Housing features such as ventilation, ceiling height, insulation, window area in modern apartment dwellings are very different from the pre-war duplex and triplex habitats and it is possible that these changes to modern dwellings may not be better in terms of disease transmission. Newer dwellings are likely to have lower ceilings, and be better insulated than older dwellings, beneficial during the harsh winters but perhaps at the detriment of better ventilation.

The variable story which indicates the number of stories of the building in which the dwelling is located, was a strong predictor. Magnitude of the effect decreased slightly when adjusted for other variables in the model. High rise buildings i.e. buildings with more than three stories are at an increased risk. A possible explanation may be lower rental value of high-rises relative to duplex and one-story dwellings, and consequent attraction to immigrants as they are more affordable. The number of stories as categorized distinguishes meaningful types of housing common in Montreal. Montreal is a city of duplexes and triplexes. 13% of buildings are 1 story, mostly likely single family and recently built, 59% are 2 story, 20% are 3 story while only 2% are taller. For 6% of buildings we had no information on the number of stories. This categorisation of buildings by stories was only possible using data from the City housing GIS.

A recent 2-year study carried out in Montreal between 1997-1998 indicated that between 82 – 96 % of incident active TB cases were due to reactivation (Kulaga et al., 2002). High-rise buildings appear to be a strong risk factor for transmission and it may be speculated that the relative infrequency of high rise buildings may be a reason that transmission of TB in Montreal is relatively infrequent.

Note that the crowding, residential density and dwelling measures indirectly provide some indication of ones socio-economic status, but as a better proxy of socio-economic status we use the variables dwelling value and median income.

Dwelling value from the housing database provided a good proxy for socio-economic status. As compared with median income from the census tract level data, it is a more refined individual-level estimate where median income is merely an ecological variable imputed from the average for the tract. The relationship between economic status, as measured by both variables and the risk of being a case of TB was as expected, more severe with lower income status. Poorer areas are at an increased risk.

As expected, the variables recent immigrants and immigrants were very strong predictor. The variable recent immigrants was a slightly stronger predictor and provides a good indication of whether an area is currently attracting new immigrants. This is particularly pertinent in that various investigators have repeatedly shown that risk is higher during the first two to five years after arrival (McCarthy, 1984, Zuber et al. 1997).

Even though localized crowding (floor area per person) was did not turn out to be a useful measure in describing TB occurrence, the residential density measures were useful and were also obtained from the block level. Since crowding is not be uniformly distributed and appears to vary widely from neighbourhood to neighbourhood we believe that our index helps us to better understand the effects of localized pockets of crowding (figure 12). Crowding estimates from the census tract level appear to play an important role in describing TB occurrence, however when crowding estimates were examined at the city block level (floor area per person) the effect of crowding on explaining TB occurrence was not significant.

Another highlight of the study was the use of a multi-level approach with use of both ecological (census tract level) and individual dwelling level data. We attempted to improve on previous studies that only looked at census measures of crowding and socio-economic status. Individual level data such as dwelling value was used to provide a complementary measure of socio-economic status, while floor area per person calculated from the city block level was used to provide a complementary crowding measure at a finer scale. The use of individual level data was absolutely necessary as it minimises errors that may be attributable to the ecologic fallacy.

Note that trying to obtain a similar socio-economic status index from a finer scale using dwelling value may have its own limitations including the fact that even though a dwelling has a very low market value it does not necessarily mean that it is run down/poorly ventilated or more likely to be at higher risk for TB. However, we believe that the dwelling sale value provides a good proxy for rental value, and in agreement with our initial reasoning, dwellings with lower market value were at higher risk. This may be because they also have a low rental value and therefore attract more people who could not afford to pay high rent. An improvement of this possible weakness would be to obtain the actual dwelling rental value as an indicator of the socio-economic status of the dwelling habitants. If the dwelling rental value could be obtained from the housing and planning department these values could possibly be linked using a GIS to the actual dwellings. For our study the rental values were not available.

Other strengths of the study included the use of (1) a relatively new tool (GIS) to assist in the linkage of datasets in an efficient and effective manner, thus enabling us to carry out a traditional case-control approach and (2) the use of the City of Montreal housing geo-database. Note that without a detailed housing GIS we would have been unable to evaluate several of the housing variables without undertaking a very time consuming survey process to collect the relevant information. The value of individual level data was demonstrated through results shown in the study.

The GIS software was essential as it enabled us to (1) geocode our cases precisely and thus enable attribution of geographically referenced data and aids in issues of spatial aggregation (2) link data from several sources and from different levels of scale (census data, individual level housing data) (3) visualise our data by any of the available variables in the dataset. In addition to geocoding, cases could easily be examined visually in relation to any census or housing variables that we decided to map. In the examples presented (figures 6, 7, 12) it is easy to visualise our data. This alone is a powerful way of justifying targeted programs that could provide more awareness in respect to availability of treatment options, access to health care facilities etc.

The City housing GIS facilitated the precise mapping of all active TB cases in the old municipality of Montreal by their residential address, as well as provided the individual dwelling attributes. It also provided the base from which control dwellings were selected.

The control dwellings were randomly sampled from the list of all residential dwellings in the City housing database. They are a representative sample of the distribution of potential exposures from which the cases arose as they are obtained from the same reference population as our cases. The study of case and control dwellings as opposed to people was an innovative and valid approach.

There are a several possible sources of bias in the methods, but the effects should not invalidate the basic findings of the multivariate model. Limitations of census data are fairly well understood, and we need to question more carefully the municipal housing source for which we have little documentation. Such sources exist in most cities, but their precision, up-to-date status and design may vary. Additionally both data sources are administrative in nature and were not specifically designed for this type of research.

Because some census tracts enclose large non-residential areas such as industrial parks, regional parks, or the airport, the population density within liveable areas of these tracts is higher than the density averaged over the entire tract. This may cause a slight bias of the results towards the null. However the study was restricted to the former City of Montreal which in fact excluded the airport and most industrial parks. Therefore this potential limitation does not affect our results.

Control dwellings were randomly selected from the total pool of residential dwellings in 2000 and therefore no bias should be involved in the control dwelling sample. However, the case dwellings that were included into the study were those that appeared in the most current (2000) version of the City housing GIS. The loss of information due to the 24 demolished dwellings may be the source of a selection bias for case dwellings. Since the numbers are small it should not bias the results significantly.

There may have been errors in the measurements of some of the variables obtained from the Canadian Census and the City housing data, however, any errors in these measurements should be non-differential i.e. they affect both cases and controls.

One constraint of the City housing database was the absence of comparable housing information for the 28 suburban municipalities that were not part of Montreal until the amalgamation of January 1st 2002. Therefore analyses were restricted to cases within the City of Montreal. The City, as administratively defined in the 1990s, accounted for 57% of the population of the Island and 72% of the TB caseload during 1996-2000.

The rate at which people change addresses in Montreal is estimated to be once every seven years and for populations at high risk of TB such as newly arrived refugees this is expected to be much higher. Given the dynamic nature of household membership especially among this population, we may inappropriately attribute data based on the time of reporting rather than time at risk and may miss transmission associated with a previous address. However, Montreal residents especially tenant households, are well known for short distance moves (within the same parish), (Germain, 2001). As well, people in a given income bracket may tend to move to another dwelling within their budget and therefore with similar locality features.

Our measure of crowding at the census tract level (people per room) was obtained by dividing an average estimate of the number of people in a household at the census tract level by an average estimate of the number of rooms in a household at the census tract level and therefore this estimate of crowding may not be very meaningful, however it appears to be providing reasonable estimates in the study.

The variable floor area per person was based on estimates of the number of people in a city block from the postal code. The postal codes used for an estimation of the population were only those reported by census respondents. Therefore, they do not constitute all valid postal codes in Canada according to Canada Post Corporation, at the time of the 1996 Census (32). This may result in an overestimate of the floor area per person, but since the underestimates of the population are expected to be to be non differential, (the same for both case and control blocks) the conclusions drawn remain the same.

Summary and conclusion

A high precision geographic information system complemented census data in pinpointing and characterizing the occurrence of TB in Montreal. It provided a more refined assessment of the impact of local crowding, after adjustment for other important factors as well as provided new information of dwelling level risk factors. Extremes of poverty and overcrowding were rare although the City housing database did enable us to identify pockets of crowded city blocks (figure 12). The precise mapping of all incident TB cases for the period 1996-2000 in the former City of Montreal was also achieved.

Limitations of aggregate data must be acknowledged, however several individual level housing risk factors were independently associated with TB occurrence. These were younger dwelling age, lower dwelling value and the higher number of stories. They are indirect indicators of socio-economic status and as expected indicate that poorer neighbourhoods tend to be at a higher risk. Note however, that these indicators may also indicate building issues such as ventilation. Contrary to expectations, localized room crowding as measured by floor area per person was not found to be a significant risk factor, but the variable lot coverage which provided a measure of residential density indicated that transmission may be due to increased encounters.

This study has shown the usefulness of the City housing database and its potential use in future studies for mapping cases, for finding the shortest distance to health facilities, for an analysis of possible shortages in health services in certain areas of the city. From a public health perspective it has the potential to be aid in decision making. The randomly distributed control dwellings could also be used to study other diseases using a similar case cohort approach. Other studies on TB could correlate spatial distances between TB patients' home addresses with genetic similarity between the infecting organisms to provide an indication of the probability of person-to-person transmission.

References

1. Acevedo-Garcia D. Zip-code level risk factors for tuberculosis neighbourhood environment and residential segregation in New Jersey, 1985-1992. Am. J. Public Health 2001; 91 734-741.
2. Amrhein CG, Reynolds H. Using a spatial dataset generator in an empirical analyses of aggregation effects on univariate statistics. 1997; Geographical and Environmental Modeling 1 199-219.
3. Bailey Trevor C, Gatrell AC. 1995. Interactive spatial Data Analyses; Longman Scientific and Technical.
4. Barr RG, Diez-Roux AV, Knirsch CA, Pablos-Méndez A. Neighborhood poverty and the resurgence of tuberculosis in New York City, 1984-1992. . Am. J. Public Health 2001; 91 1487-1493. [Full text]
5. Beyers, N.; Gie, R.P.; Zietsman, H.L.; Kunneke, M.; Hauman, J.; Tatley, M.; Donald, P.R. 1996. The use of a geographical information system (GIS) to evaluate the distribution tuberculosis in a high-incidence community. *S Afr Med J.* 8640-44.
6. Bugayevskiy LM, Snyder JP. 1995. Map Projections A Reference Manual. London Taylor and Francis.
7. Bureau de surveillance épidémiologique, Direction régionale de la santé publique de Montreal. Tuberculose, Région de Montreal-Centre, 1995-1998
8. Cantwell MF, McKenna MT, McCray E, Onorato IM. Tuberculosis and race/ethnicity in the United States impact of socio-economic status. Am. J. Respir Crit Care Med 1997; 157 1016-1020. [Full Text]
9. Comstock, G. W., and G. M. Cauthen. 1992. Epidemiology of tuberculosis.In L. B. Reichman and E. Hershfield, editors. Tuberculosis a Comprehensive International Approach. Marcel Dekker, New York. 23–48.
10. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. J R Statist. Soc. 1990; Series B 52 73-104.
11. de Savigny, D., Loslier L., Chauvin J., Synergy – GIS for Health and the Environment Foreword, IDRC 1995
12. Dennis L, Drga L, Gantz Donald T. Combining Statistical and Spatial Analyses to characterize Tuberculosis Incidence in Northern Virginia 1989-2000. Paper

presented at the NCHS Cartography and GIS lecture Series, October 2001

13. Dennis L, Drga L, Gantz Donald T. Ten years Data of Active Tuberculosis in Fairfax County, Virginia. IDR 2000;2(suppl)37.
14. Dennis L, Drga L, Gantz Donald T. Using GIS Maps to Visualize the Relationship between TB Incidence and Socio-economic Factors and to Identify Neighborhoods with High TB Risk. IDR 2000;2(suppl)19.
15. Dye C, Scheele S, Dolin P, et al. *Global burden of tuberculosis. Estimated incidence, prevalence, and mortality by country.* JAMA 1999;282677-86.
16. El Sahly, H.M.; Adams, G.J.; Soini, H.;Teeter, L.;Musser, J.M.;Graviss, E.A. 2001. Epidemiologic differences between United States- and recent immigrants tuberculosis patients in Houston, Texas. *The Journal of Infectious Diseases.* 183 461-8.
17. Elender F, Bentham G, Langford I. Tuberculosis mortality in England and Wales during 1982-1992 its association with poverty, ethnicity and AIDS. Soc Sci Med. 1998; 46673-681.
18. Fanning A. E. 1998. Globalization of tuberculosis. CMAJ 1998; 158 (5) 611-612.
19. Fotheringham S, Charlton M, Brunsdon C. Spatial Variations in the Relationship between 'Health' and 'Wealth' across England and Wales. Department of Geography, University of Newcastle, UK. Paper presented at the 2nd International Health Geographics Conference National 4-H Conference Center Chevy Chase, MD, March 2000
20. Germain, Annick (2001). Les métropoles comme pôles de circulation et lieux de rencontre : la dynamique de Montreal, communication au colloque Les grands récits des Amériques : polyphonie des identités culturelles dans le contexte de la continentalisation des Amériques, Montreal, 10 avril, 6 p.
21. Guerra G. Mcgill University. March 2001
22. Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N. Childhood leukemia in Sweden Using GIS and a spatial scan statistic for cluster detection. Statistics in Medicine, 1996;15707-715.
23. Hosmer DW, Lemeshow S. Applied Logistic Regression, 2nd ed. 2000; John Wiley & Sons, Inc.
24. Hwang, SW. 2001 Homelessness and health. CMAJ January 2001; 164(2)
25. Iskander, R. TB control in London-The next steps, Project progress report of

the London TB group, March 2001.

26. Kamel, W. W. (1997) Health dilemmas at the borders-a global challenge. World Health Forum, 18, 9-16.

27. Kazmi, J.H., Pandit K. Disease and dislocation the impact of refugee movements on the geography of malaria in NWFP, Pakistan. Social Science and Medicine 52 (2001) 1043-1055.

28. Kleeberg HH, Olivier MS. A world Atlas of Initial Drug Resistance. 2nd ed. Pretoria Tuberculosis Research Institute of South African Medical Research Council, 1984.

29. Kleinschmidt, I.; Bagayoko, M.; Clarke, G.P.Y.; Craig, M.; Le Sueur, D. 2000. A spatial statistical approach to malaria mapping. *International Journal of Epidemiology*. 29355-361.

30. Kulaga, S.; Behr, M.; Musana, K.; Brinkman, J.; Menzies, D.; Brassard, P.; Kunimoto, D.; Tannenbaum, T.; Thibert, L.; Joseph, J.; Boivin, J.B.; Schwartzman, K. Molecular epidemiology of tuberculosis in Montreal. CMAJ August 2002; 167 (4)

31. Kulldorff M, Nagarwalla N. Spatial disease clusters Detection and Inference. Statistics in Medicine, 1995;14799-810.

32. Kulldorff M, Rand K, Gherman G, et al. SaTScan version 2.1.3. Software for spatial and space-time statistics. National Cancer Institute, Bethesda, MD, 1998.

33. Lee J, Wong David WS. Statistical Analyses with ArcView GIS. 2001 John Wiley & Sons, Inc.

34. Lerner BH. New York City's tuberculosis control efforts the historical limitations of the "war on consumption". Am J Public Health. 1993;83758-766

35. Long R, Njoo H, Hershfield E. Tuberculosis 3. Epidemiology of the disease in Canada. CMAJ 1999; 160 1185-90.

36. Long R. 2000. Canadian Tuberculosis standards. 5th Edition. Canadian Lung Association, Canadian Thoracic Society and Health Canada

37. Mangtani P, Jolley DJ, Watson JM, Rodrigues LC. Socio-economic deprivation and notification rates for tuberculosis in London during 1982-91. BMJ. 1995;310(6985)963-966

38. McCarthy, OR. Asian immigrant tuberculosis-The effect of visiting Asia. BR. J. Dis. Chest (1984) 78248

39. McMurray DN, Kimball MS, Tetzlaff CL, Mintzer CL. Effects of protein deprivation and BCG vaccination on alveolar macrophage function in pulmonary tuberculosis. *Am Rev Respir Dis.* 1986;1331081–1085.[Medline]

40. Moore, DA., Carpenter TE. Spatial analytical methods and Geographic Information Systems Use in Health Research and Epidemiology. 1999 Epidemiologic Reviews.

41. Njoo H. *Tuberculosis—a re-emerging public health threat in Canada.* Can. J. Infect. Dis. 1998;9(5)273-5.

42. Pensylvania West nile surveillance program. 2002 Http://www.westnile.state.pa.us/

43. Prentice, RL. A Case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986. 731-11

44. Prothero, R.M. Disease and mobility A neglected factor in epidemiology. (1977) International Journal of Epidemiology, 6(3), 259-267.

45. Reichman LB, Hershfield ES. Tuberculosis A Comprehensive International Approach, 2nd ed. 2000 144 217-221

46. Riley RL, Mills CC, O'Grady, et al. Infectiousness of air from a tuberculosis ward. Am. Rev. Respir. Dis. 1962;85511-525.

47. Rivest P, Bédard L, Tannenbaum T. Epidémiologie de la tuberculose Région de Montreal-Centre, 1995-98. Montreal Direction de la santé publique de Montreal-Centre, 1999.

48. Rivest P, Tannenbaum T, Bédard L. Epidemiology of tuberculosis in Montreal. CMAJ Mar. 10, 1998; 158 (5)

49. Ross NA, Nobrega K, Dunn JR. Income inequality, income segregation and mortality in North American cities. Submitted, Geojournal, 2001.

50. Rothman KJ. and Greenland S. 1998 Modern Epidemiology Second edition

51. Rusen, I. D., Yuan, L., Millson, M. E. Prevalence of Mycobacterium tuberculosis infection among injection drug users in Toronto CMAJ 1999 160 799-802

52. Scholten HJ, de Lepper MJC. 1991. The benefits of the application of geographical information systems in public and environmental health. WHO Statistical Quarterly, 44(3).

53. Snow, SJ, 2002. Commentary: Sutherland, Snow and water: the transmission of cholera in the nineteenth century. International Journal of Epidemiology

2002;31:908-911

54. Spence DP, Hotchkiss J, Williams CS, Davies PD. Tuberculosis and poverty. BMJ. 1993;307 (6907)759–761.
55. Statistics Canada. 1996 census dictionary—final edition. 2001 Catalogue No. 92-351-UIE. (Available at www.statcan.ca)
56. Statistics Canada. 1996. Census of Canada, Cumulative profiles No. 95F0183XDB
57. Statistics Canada. 2000. Postal Code Conversion File - Reference Guide. Catalogue No 92F0027XDB
58. Statistics Canada. 1996. Postal Code Counts - Population and Dwelling Counts. Catalogue No. 92-F0086-XCB
59. Stone LM. The utility of Geographical information systems (GIS) and Spatial analyses in Tuberculosis Surveillance in Harris County, Texas, 1995-1998. (2001) Public Health and GIS Researcher, Center for Health Policy Studies. University of Texas-Houston School of Public Health
60. Strachan DP, Powell KJ, Thaker A, Millard FJ, Maxwell JD. Vegetarian diet as a risk factor for tuberculosis in immigrant south London Asians. *Thorax.* 1995;50175–180.[Abstract]
61. Szwarcwald Célia L, Tavares de Andrade, Carla Lourenço. Geographic Pattern of Poverty in Rio de Janeiro, Brazil. Fundação Oswaldo Cruz, Rio de Janeiro, Brazil. Paper presented at the 2nd International Health Geographics Conference National 4-H Conference Center Chevy Chase, MD, March 2000
62. Tim S. The Application of GIS in Environmental Health Sciences Opportunities and Limitations Environmental Research Vol. 71, No. 2, November 1, 1995
63. Tuberculosis Fact Sheet No 104. Information Resource Centre, Communicable diseases, WHO 2002.World Health Organization. Global Tuberculosis Control Surveillance, Planning, Financing. WHO Report 2002. Geneva, Switzerland, WHO/CDS/TB/2002.295
64. Turnbull B, Iwano EJ, Burnett WS et al. Monitoring for clusters of disease application to leukemia incidence in upstate New York. Am. J. Epidemiol. 1990 132.
65. Using GIS to Identify Gaps in Health Services. Errol Bamford. IDR 2000;2(suppl)21.

66. Ville de Montreal. Géomatique et développement des systèmes SIURS. www.ville.Montreal.qc.ca/urb_demo/domaines/geomatique/siurs.htm

67. Vine M, Degnan D, Hanchette C. Geographic Information Systems Their use in Environmental epidemiological Research. October 1998. Environmental health

68. Wessen, A.F. (1974) The role of migrant studies in epidemiological literature. Israel Journal of Medical Science, 1, 584.

69. World Development Report 1993 Investing in Health. Oxford Oxford University Press, 1993.

70. World Development Report 1993 Investing in Health. Oxford. Oxford University Press, 1993.

71. Zuber, PL, McKenna, MT, Binkin, NJ, Onorato, IM, Castro, KG. Long term risk of tuberculosis among recent immigrants persons in the United States. JAMA 1997. Vol. 278. No. 4.

Figures:

Figure 1: Reported TB incidence and mortality in Canada, 1924 – 1998

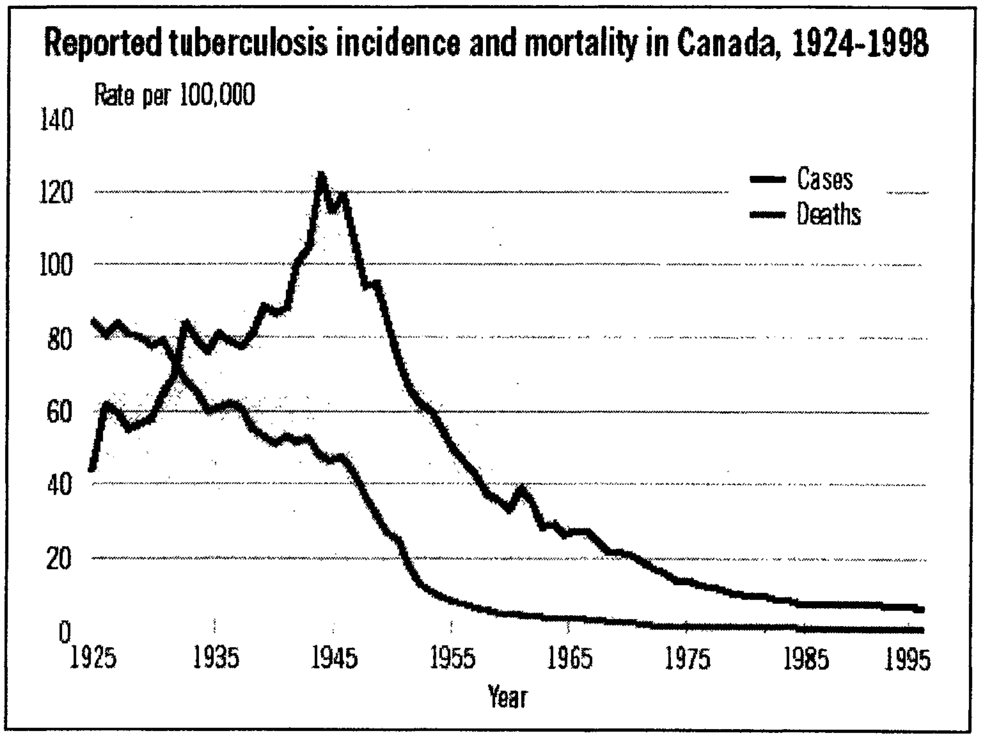


Figure 2: Montreal census tracts by median income (1996 data)

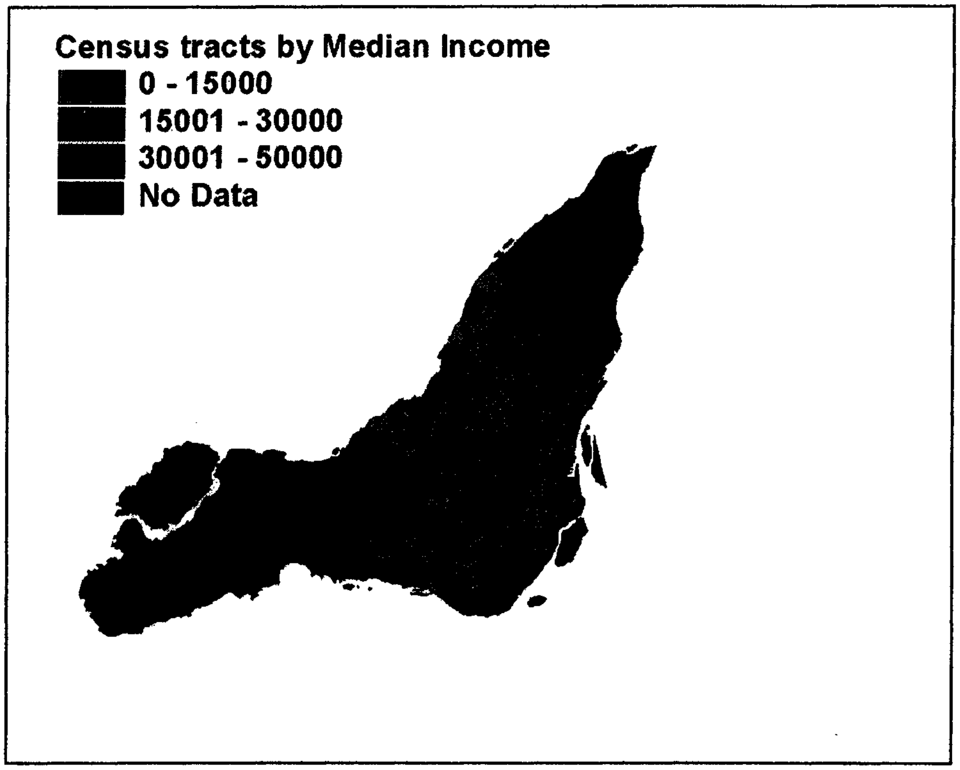


Figure 3: Immigrant TB cases in Canada (Long, 2000)

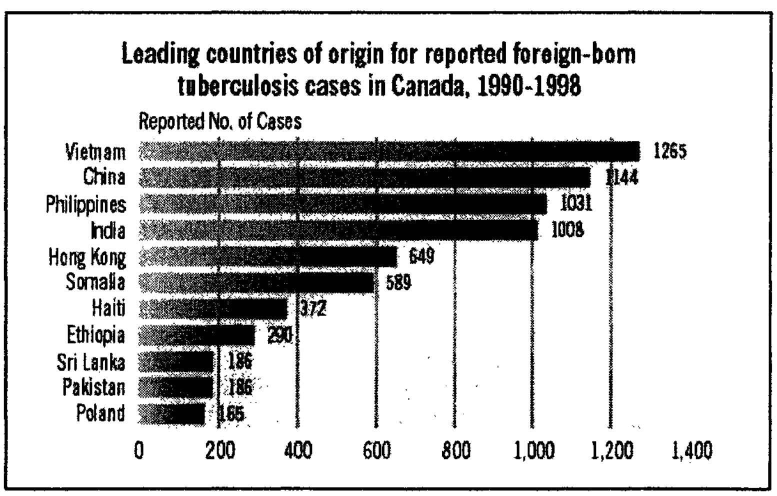


Figure 4: TB incidence in Canada by area of origin

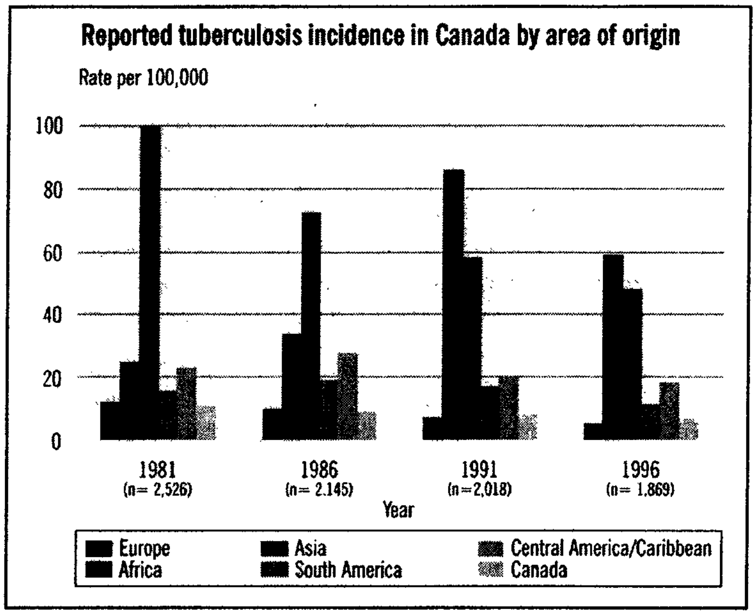


Figure 5: GIS data structure

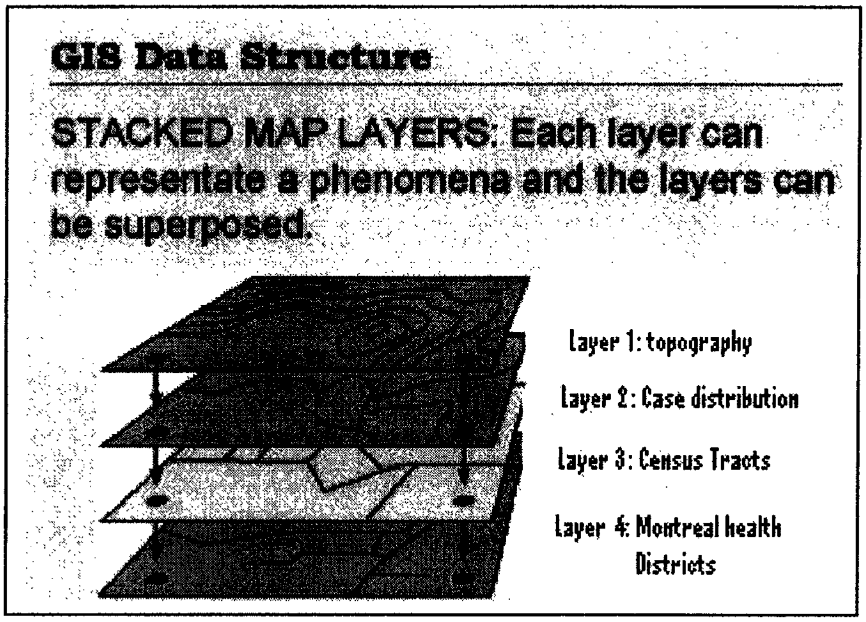
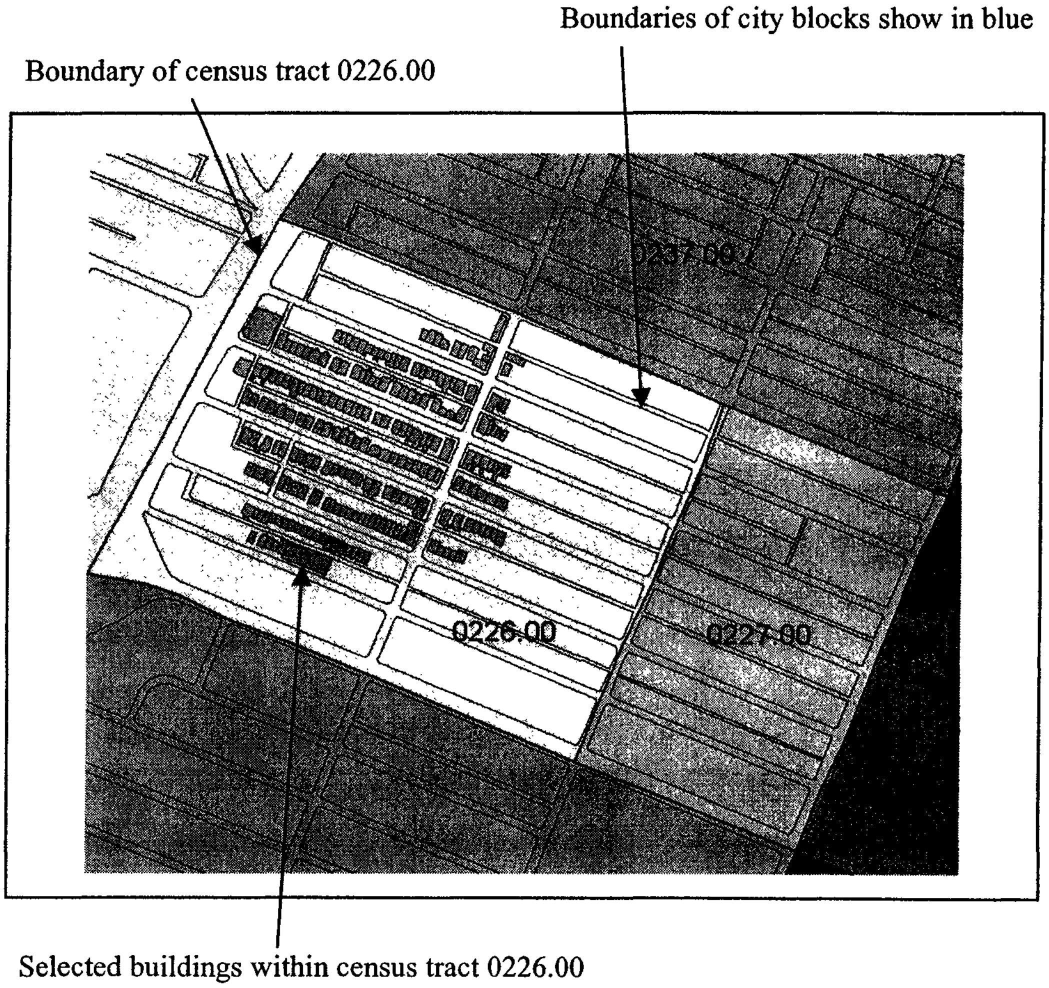


Figure 6: Selected census tracts, city blocks and buildings[5]



[5] Building refers to a housing structure, which may contain a single or multiple units within it. Dwelling for the purpose of this thesis refers to a single unit in which people reside (house or apartment unit within an apartment building or a single unit within a duplex/triplex/etc).

Figure 7: Montreal TB cases (1996-2000) by municipality

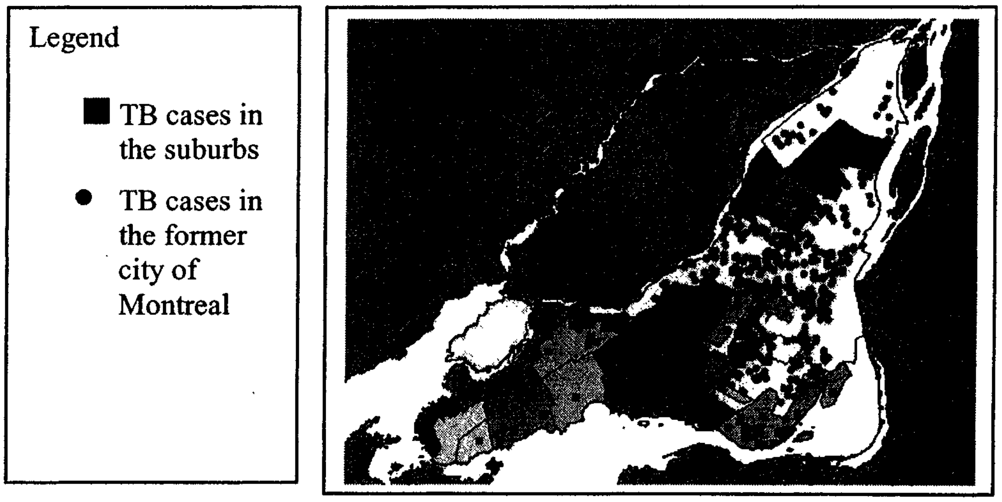


Figure 8: Floor area per person calculation

1) Total floor area for dwellings within selected block is

(200 x 2) + (220 x 5) = 1500 m^2

2) Total population of that block, obtained from the population count of the six-digit postal code(s) within the block = 35

3) Floor area per person = 1500 / 35 = 42.8 m^2 per person

4) All dwellings within this block will be assigned this estimate of floor area/person

In this case only one six-digit postal code corresponds to the city block (population=35 people)

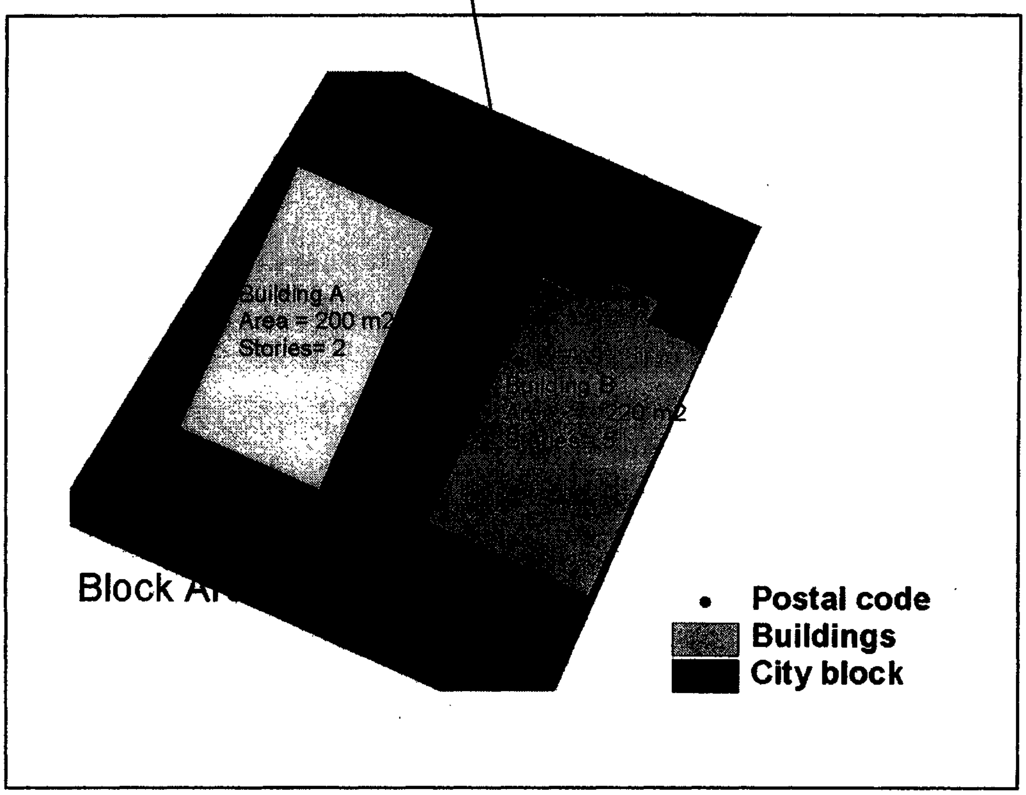


Figure 9: Building age distribution

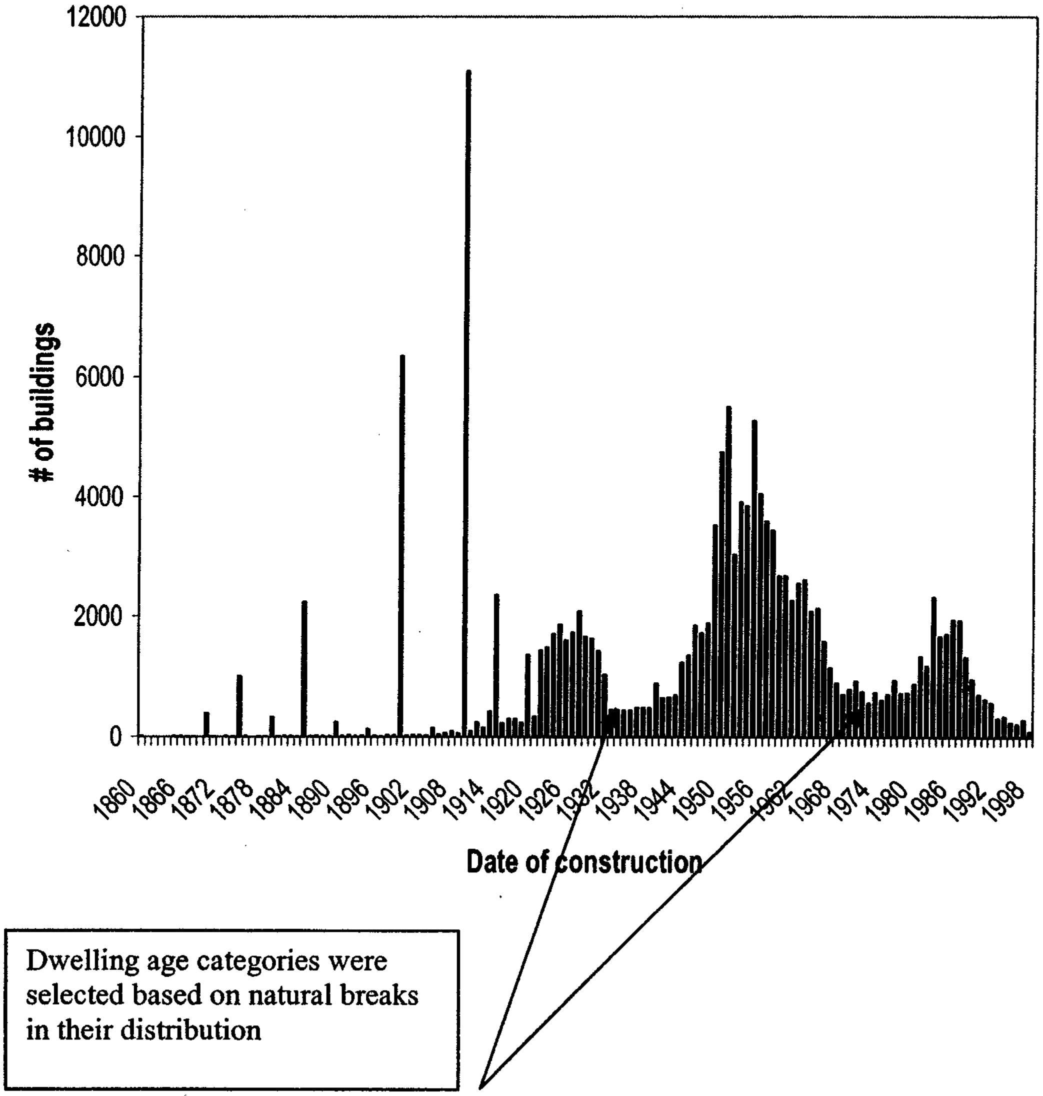


Figure 10: Graph of the unadjusted odds ratios with the 95% CI

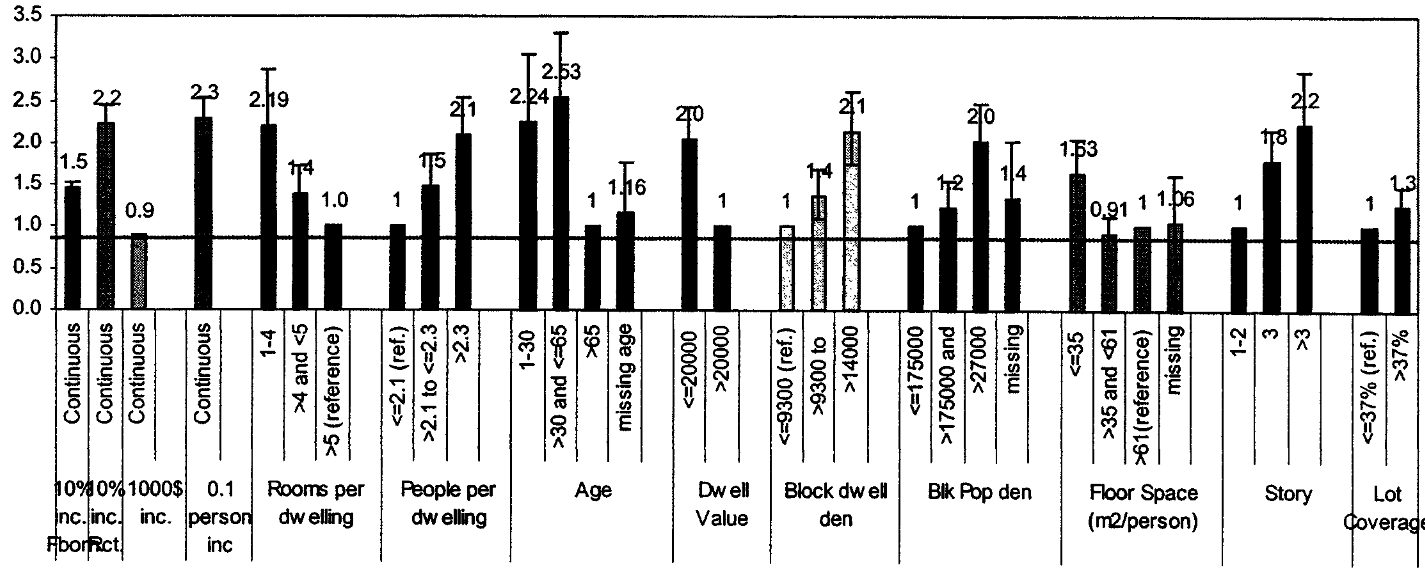


Figure 11: Comparison of the Crude and Adjusted odds ratio

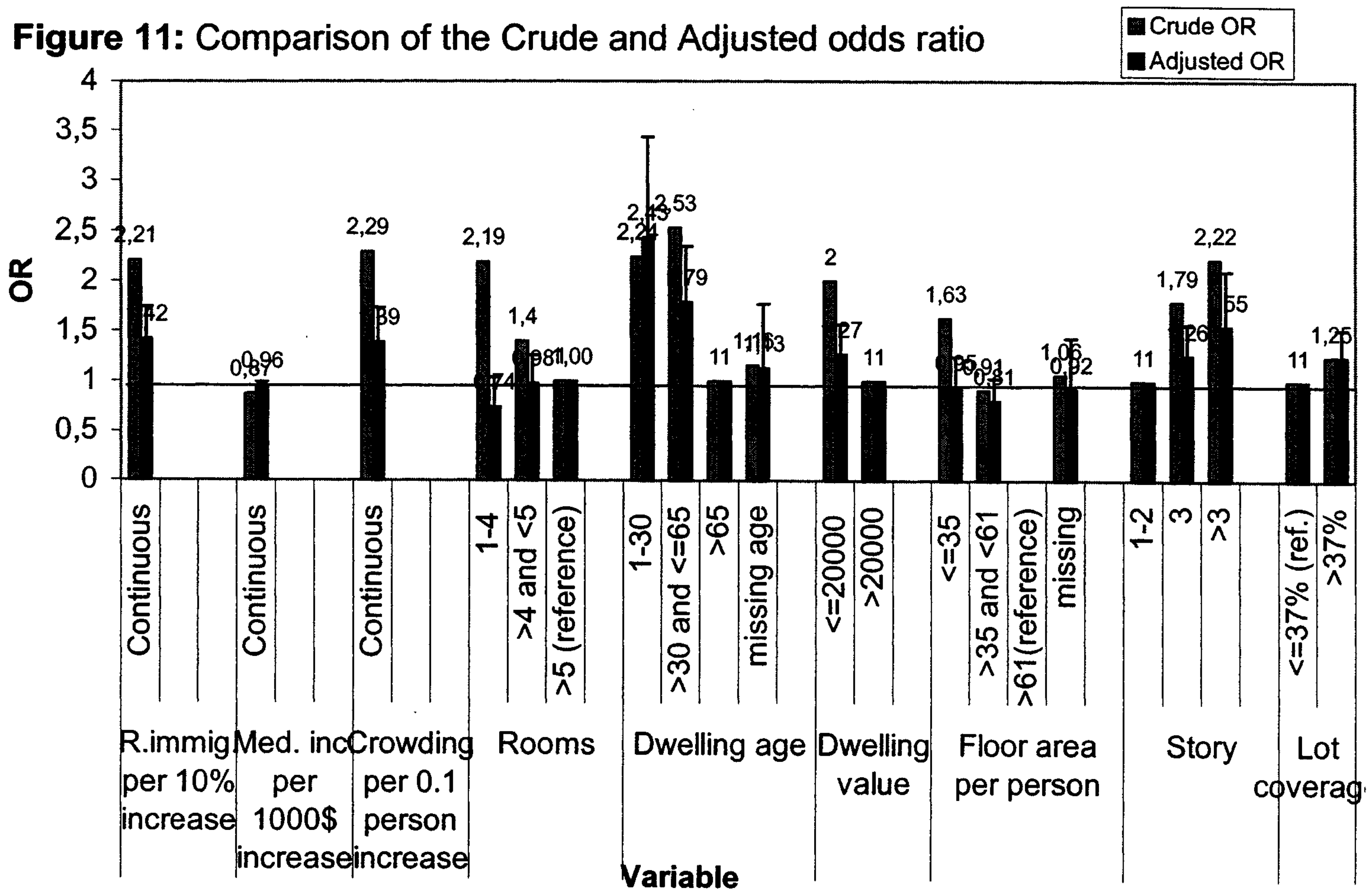
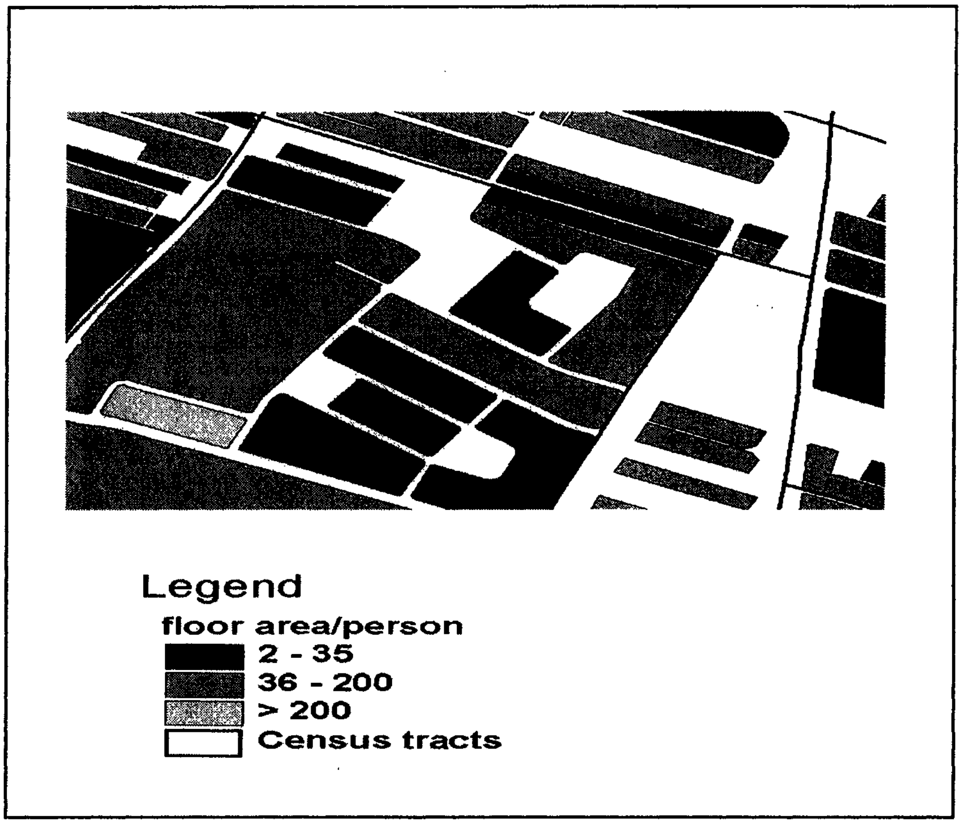
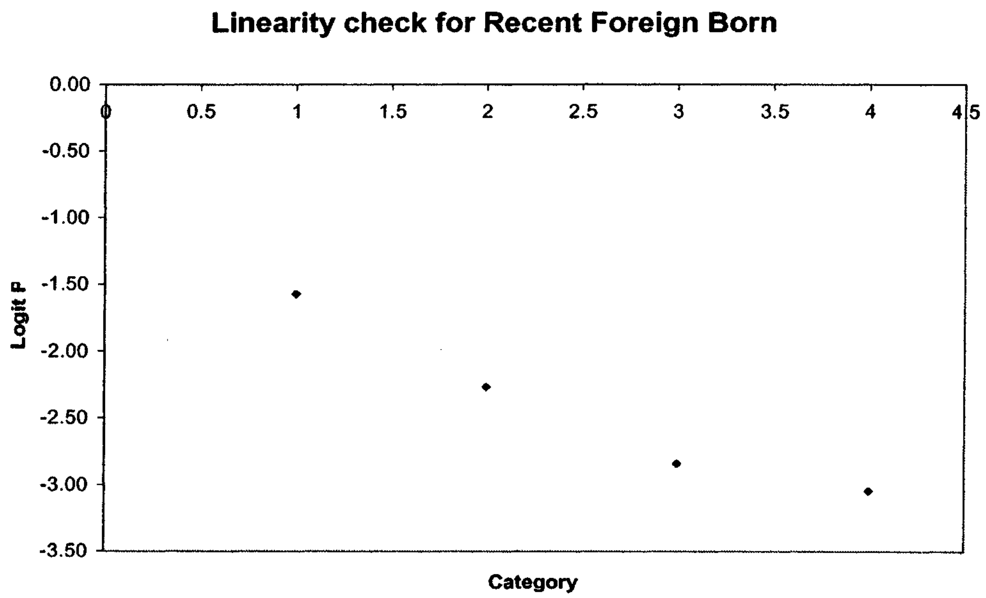


Figure 12: Example of localized crowding

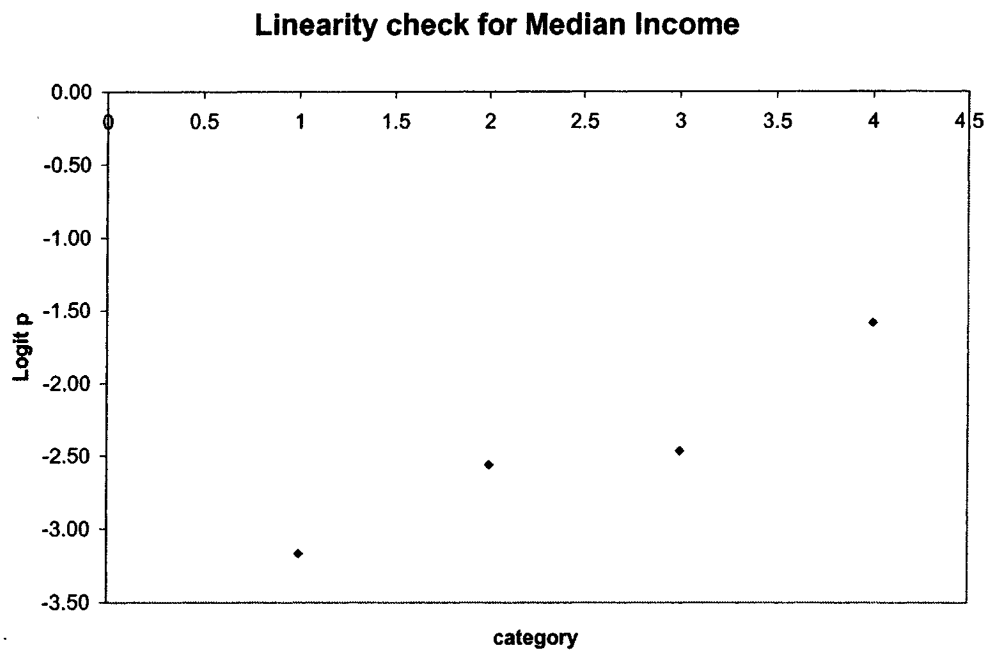


Appendices

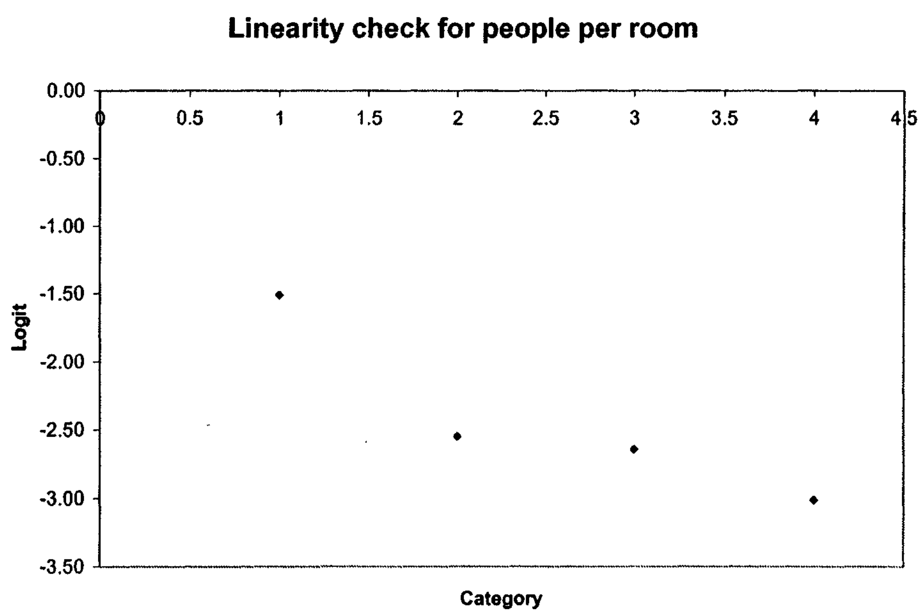
Appendix 1: Linearity check for recent immigrants



Appendix 2: Linearity check for median income



Appendix 3: Linearity check for people per room



Appendix 4: Detailed control selection for the City of Montreal

Each TB case is assigned to a building[5]. Each building may consist of several dwellings. In order for us to obtain a representative sample of the distribution of dwellings, each building must be weighted by the number of dwellings it contains. Our unit of analyses is thus considered to be a dwelling.

In order for our controls to be a representative sample of the dwelling distribution in Montreal, all residential buildings are enumerated and a random sample of 5950 dwellings are selected from the total population of residential dwellings.

There are 468,130 residential dwellings in the City of Montreal (determined by obtaining the sum of dwellings for all residential buildings from the SIURS).

Each of the 468,130 residential dwellings is assigned a unique number. A list with 468,130 numbers is generated and each of these numbers is assigned a random number (generated using SAS code).

This list is then ordered by the random number from the lowest to highest.

There are 595 cases that occur within the City. Since we are sampling from the base of dwellings in the ratio of 1:10 (Case dwelling:control dwelling) we require 5950 control dwellings. Therefore the first 5950 numbers are selected from the dwelling list.

To obtain the records that correspond to the 5950 randomly selected control dwellings, a relational database is used to match the 5950 randomly selected numbers to the appropriate control dwelling.

Appendix 5: Categorization of variables

Using the variable dwelling value as an example. There are 595 cases and 5950 controls. The distribution of the control set is obtained and the quartiles are found to be as follows

75% Q3 49700.000

50% Median 30766.667

25% Q1 20210.000

The variable is then categorized based on these quartiles as follows

if dwellv <=20000 then dwellcat=1;

else if dwellv >20000 and dwellv <=30000 then dwellcat=2;

else if dwellv >30000 and dwellv <=50000 then dwellcat=3; else dwellcat=4;

Following this the crude odds ratio pattern can be obtained. This enables us to examine whether the variable can be used as a continuous variable as well as determine whether so many categories are useful

Frequency	0	1	Total
(Cheapest) 1	1456	237	1693
2	1440	118	1558
3	1587	131	1718
(Most expensive)4	1467	109	1576
total	5950	595	6545

The odds ratios obtained in relation to category 1 (least expensive dwellings are obtained as follows

118x 1456/(1440*237) = 0.50

(131*1456)/(1587*237) = 0.51

(109*1456) / 1467*237) = 0.46

Since all categories are in the same direction i.e. all at decreased risk relative to category one it is not necessary to have so many categories and a decision is made to have only two categories; one with dwellings below 20000 and the other with dwellings above 20000.