DEVELOPMENT OF BASEMAP FOR LONGITUDINAL DATABASES

SILVANE PAIXÃO

RAAFEY MOHAMMED

JOHN STEEVES

Dalhousie Medicine New Brunswick - DMNB
Faculty of Medicine, Saint John, Canada
s.paixao@dal.ca, john.steeves@dal.ca

University of New Brunswick - UNB
Computer Science Department, Saint John, Canada
raafey.mohammed@unb.ca

RESUMO – SIG (Sistemas de Informação Geográfica) tem sido uma ferramenta essencial utilizada para averiguar as relações espaciais entre áreas aplicadas e a saúde pública. Distribuição espacial de uma doença, a distribuição do acesso à saúde por parte da população ou da escassez de profissionais da área de saúde são apenas alguns exemplos de aplicações existentes. Dalhousie University - Faculdade de Medicina (Canadá) iniciou um projeto para identificar a localização de seus graduados em medicina e sua respectiva escolha profissional. Outro objetivo desta pesquisa foi o de investigar se existe alguma característica demográfica (por exemplo, origens em áreas rurais) ou currículo de formação médica que tenha um impacto sobre suas carreiras. Para alcançar esse objetivo, uma aplicação de SIG foi desenvolvida com objetivo de criar um mapa de referência com dados longitudinais. Critérios de integração entre os dados estáticos-históricos dos estudantes e os dados dinâmicos de população/categorização das redes urbanas e rurais também foram modelados. Este trabalho aborda o desenvolvimento de uma metodologia para criar um mapa de referência para projetos longitudinais, o mesmo envolve várias etapas de geoprocessamento e representação espacial.

ABSTRACT - GIS (Geographical Information Systems) is becoming an essential tool utilized to investigate the spatial relationships between area context and public health. Geospatial distribution of a disease, distribution of the access of health care by the population or primary care workforce representing shortage areas for health professionals are just some examples of existing applications. Dalhousie University – Faculty of Medicine (Canada) initiated a project to not only identify the location of its medical graduates and their career choice, but also investigate if is there any demographics characteristics (such as been originally from rural areas) or medical training curriculum (such as been exposed to rural rotations during medical training) will impact on their health professional careers. In order to reach this goal, a GIS application was developed to create a basemap with longitudinal data. In additional, criteria were creating linking static-historical data with dynamic data of the population (community size). This paper addresses the development of a methodology to create a basemap for longitudinal projects which involved multiple geoprocessing steps and spatial representation.

1 INTRODUCTION

Over the years GIS has been successfully applied in many areas of public health ranging from assessing health facility placement to creating communicable disease early warning systems, to describing the spatial and temporal pattern of disease incidence and to relate them to environmental and socio-economic causal factors (HIGGS; GOULD, 2001).

As Hart et al. (2005) commented, GIS application is an important tool for the development of targeted interventions leading to the following better health outcomes and/or reductions in the cost of service provision respectively: by defining the prevalence area of diseases; by helping target a culturally-sensitive population; by indicating areas of unequal health care; by directing vector control resource to the highest priority response areas; by locating clusters in space and time and suggesting intervention sites among others. By informing epidemiologists, policy and decision makers, and health workers of the location and geographic relationship between datasets, GIS helps target existing interventions to improve the efficacy of the service delivered, or reduce costs associated with service delivery.

For human resource planning, GIS has been supporting decision makers by identifying the physician maldistribution (i.e., surplus or a shortage of the type of physicians needed to maintain the health status of a given population at an optimum level) which impacts the healthcare costs, quality, and health access to the population (DRAKE, 2009; PHILLIP JR ET AL., 2000; HIXON ET AL., 2012).

The per capita distribution of physicians is highest in urban areas leading rural and remote areas underserved in terms of medical care. The preference to urban areas is linked to the professional autonomy and lifestyle, access to technology and other resources, diverse medical cases among others (PONG ; PITBLADO, 2005; DRAKE, 2009; PITBLADO; PONG, 1999). Pitblado and Pong (1999) discussed that one of the problems faced in Canada is the inconsistent definition of "Rural" that can occur even in between departments in the same organization. Methodology to define "Rural" areas can be affected by characteristics such as physician availability, distance time to the hospital in good weather, postal code classification, population census classification and so on.

The issue of chronic maldistribution of physicians in Canada drives forces in assessing the spatial location of Dalhousie University medical graduates. Background experiences, rotation placement during the medical training and their professional practices locations are some of the spatial components. One of the long term outcomes of this program evaluation is to verify if students are practicing in the nearby provinces of both main and regional campuses (the Maritimes) and/or in rural settlements. The GIS development has been playing an important visual role in the project, along with the application that deploys summarized charts and descriptive statistical analysis of the integrated data.

One of the challenges encountered during project development was how to link static-historical data (such as students' background locations) with the dynamic data (the population categorization of the communities that changes over the time). Other GIS and data challenges were identified in previous abstracts presented (PAIXAO ET AL., 2014; MOHAMMED ET AL., 2014). It was noticed during the project design that the historical data would have to refer to the census information of the community size of each student experience during the time that it happened, which may not necessarily be the same population size as it is in present time. A basemap was created to support this longitudinal characteristic of change in population categorization over the time and was used as a reference map for the geocoding of students' location experiences. The methodology of the basemap development for longitudinal projects will be the focus of this paper.

1.1 Longitudinal Database

Ander-Peciva (2005, p.4) defined Longitudinal Database as "a mass of data providing information about individuals over time" showing ***'when'*** and ***'how'*** an event happened". Some examples of longitudinal database are for example: a) land registry systems where landowner's information or land rights are in constant change; b) census information which contains demographic data with population/ other characteristics that also changes over the time. According to the author, understanding the data linkage is as important a process as the results:

- Identify the source content, organisation and quality,
- Identify how the content is represented in the database (e.g., table format, description, spatial representation),
- Define criteria and methods used in the data linkage (e.g., rules of matching criteria) for comparative studies.

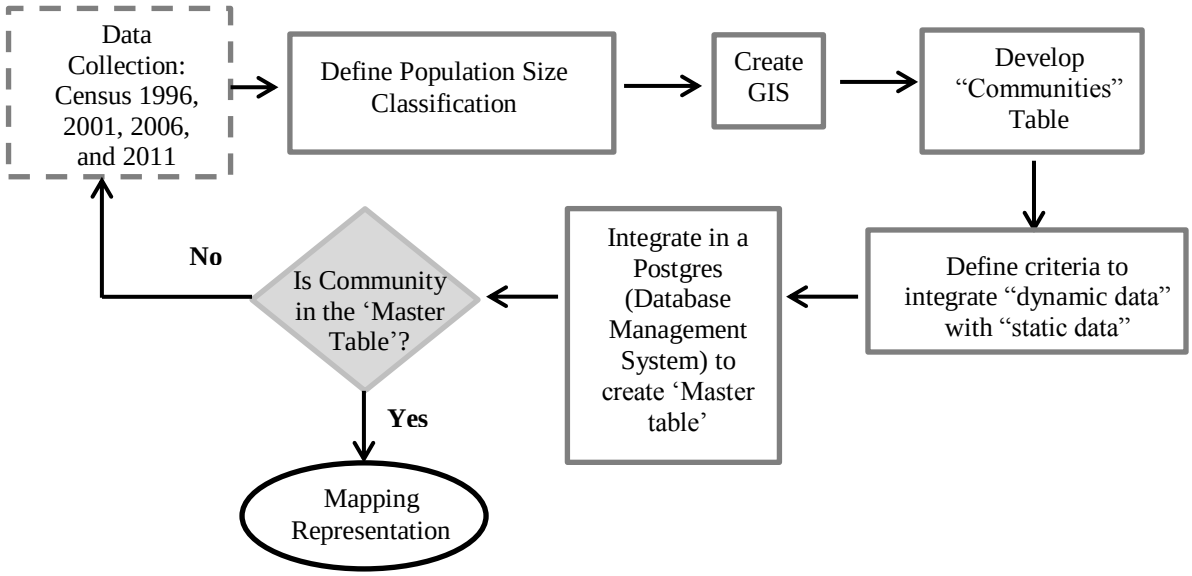
A growing number of GIS studies, have been used for monitoring public health on a longitudinal view of the influences on health. The spatial analysis has been used to evaluate "locations" and pre and post-event exposure for both longitudinal studies at the individual and aggregated geographical level or/and to track event-outcomes in subsequent years. This understanding of the relationship event-variables/ space supports the development of mitigation strategies for future exposures (BARNARD ; HU, 2005**;** CURTIS; LEITNER, 2006; SHIRAYAMA ET AL., 2009;), at the same time that GIS can also be used for near-real-time data for health planning, promotion and protection or for the identification of spatial pattern in cases over time relative to the population-at-risk (BOULOS, 2004).

1.2 LOCATED Project

A longitudinal program evaluation called the Location of Clinician and Trainee Education Dalhousie (LOCATED) Project was formed in 2010 at Dalhousie University (Dal) - Faculty of Medicine following the steps of Memorial University (MUN) in developing a tracking initiative of their learners and medical (MD) graduates. A user-friendly application linked with geographic information systems (GIS) has been designed to display/analyze integrated internal and national data (admission data/medical training experiences and career choice/practice location respectively) aiming to show the geographical distribution of career choice and practice locations of Dalhousie medical graduates

2. METHODOLOGY

In order to create the basemap to be used as reference for the longitudinal LOCATED Application, the following steps were necessary:

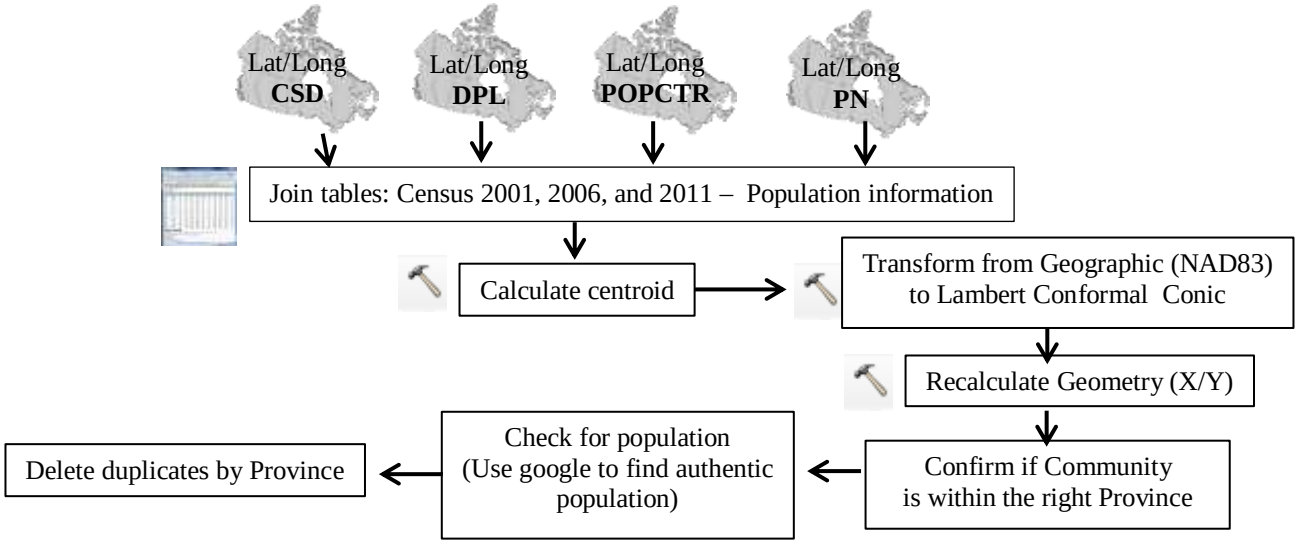


2.1 Collect Data: Census Data

In order to create the layer of communities (point) and classify them according to the Population Categories (Table 1) the following Census attribute 1996, 2001, 2006 and 2011 tables and map (polygon) transformed to centroids (points) from Statics Canada were used:

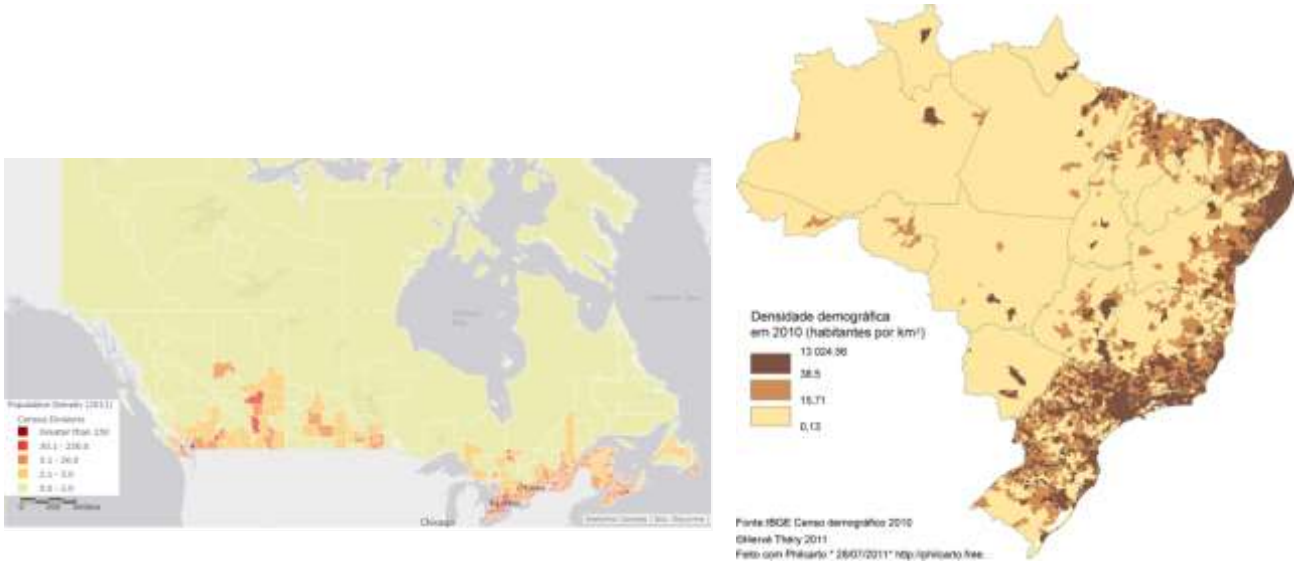
- **Census Subdivision (CSD) -** term for municipalities (as determined by provincial/territorial legislation);
- **Population centre (POPCTR) -** areas with a population of at least 1,000 and no fewer than 400 persons per square kilometre (i.e., rural areas);
- **Designated place (DPL) -** small community that does not meet the criteria to be an area with municipal status or a population centre.
- **Place Name (PN) -** Names of some local places, including CSD, DPL, PopCTR which population is not given by Statics Canada

The data clean up happened to eliminate duplicated data for communities that appear on various sources with different ID number, different Lat/Long but same population, or those with same Lat/Long but different population in the same province. Data containing same named communities in different provinces was retained. The process was:



2.2 Define Population Size Classification

There are about six Brazilians (192,376,496 estimated population – IBGE (2011)) for every Canadian (33,476,688 inhabitants). According to McCartney (2011), about 90% of Canada is uninhabited. As it can been see from Figure 1, 90% of Canadians live within 500 km of the USA border which is 8,890 km in length where largest urban centres are also found. This strip of land is about 4 percent of Canada's total area (Statistics Canada (2011)). Brazilian population is concentrated on the coast, with high population concentration at the capital cities and surrounding areas. The population density in Canada (Census 2011) is 3.7 very low compared with 22.5 people per square kilometre in Brazil.



a. b.

Figure 1 – Display Population Density in Canada (a.) and in Brazil (b.). Source: a. ESRI (2013); b. SOMAIN(2011)

In Canada, like in Brazil, the definition of the term "Rural" varies among institutions. For example, for statistical purposes, Canadians define rural population as "persons living outside centres with a population of 1,000 and outside areas with 400 persons per square kilometre" (STATISTICS CANADA, 2011). While for Brazilians, the census categorizes rural as: a) Rural Areas – area external to the urban perimeter, b) Rural Settlement - A cluster of buildings located at rural area with more than 50 housing units with population exceeding 250 inhabitants (IBGE, 2013). As short listed by Pitblado and Pong (1999) on Table 1, the variety of "Rural" definition in Canada makes difficult for researches to compare results.

Table 1 – List of criteria to define "Rural" in Canada. Source: Pitblado and Pong (1999)

Organization	Population Characteristics	Distance Characteristics	Comments
Statistics Canada	Pop <1,000 Pop Density <400 persons/sq. Km	Adjacency to Metropolitan Areas (CMA) and Census Agglomeration Areas (CA)	For each enumeration area: Urban core/ urban fringe/ rural fringe/ urban outside CMA/CA and Rural outside CMA/CA
Canadian Medical Association & Canada Post			Second digit of postal code "0"
Ontario Medical Association	Pop. <10,000	Distance to a community of 50,000	Used to define groups of physicians regarding continuing medical education subsides: Group 1 > 80km; Group 1 50-80km;
Canadian Association of Physicians	Pop. <10,000	Distance from major regional hospital: < 80km or 60min; 80-400km or 1-4h; > 400km or > 4h	Rural Close/ Rural Remote/Rural isolated

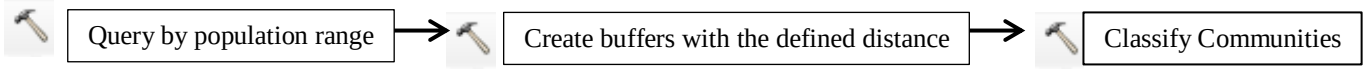
This research adopted with adjustment, the "rural" definition based upon population size and distance to an urban center developed by Memorial University (MUN) Learners and Locations: A Pilot Study of Where Physicians Train and Practice Project. The LOCATED Project decided to make adjustments because Maritimes clusters (where Dal is located) are not as distinct as Newfoundland's (where MUN is located), see Table 2.

Table 2 – Compares Population Size Categories among the two medical students tracking project

Memorial University (MUN) - Learners and Locations Project	**Dalhousie University (Dal) - LOCATED Project**
Urban	***Large Centers > 50,000***
1. Metropolis: Pop. >1,000,000 2. Very Large: 500,000 < Pop < =1,000,000 3. Large: Pop. 100,000 < Pop < = 500,000 4. Medium: 25,000 < Pop < =100,000	1. Metropolis: Pop. >1,000,000 2. Very Large: 500,000 < Pop < = 1,000,000 3. Large: Pop. 100,000< Pop < = 500,000 4. Medium: Pop. 50,000 < Pop < =100,000
Communities ***Small 10,000 – 25,000*** 5. Small Rural City: 10,000 < Pop. < = 25,000 ***Rural Communities<10,000*** 6. Small Rural Communities: Pop. <10,000	***Small Centers 10,000 -50,000*** 5. Small: 10,000 < Pop < = 50,000; <200km from larger city (1-4) 6. Small Rural: 10,000 < Pop < = 50,000; 200 – 500 km from larger city (1-4) 7. Small Remote: Pop.10,000 <Pop < = 50,000; >500km from larger city (1-4)
	Rural/ Remote Communities <10,000 8. Rural-urban: Pop. <10,000; <60km from medium or larger city (1-4) 9. Rural-close: Pop. <10,000; <60 km from small city (5 to 7) 10. Rural: Pop. <10,000; 60-200km from city 11. Remote: Pop.<10,000; >200km from city

2.3 Use of Population Size Classification at GIS

GIS Application was developed to classify the communities using the following geoprocessing steps. Classification was done from large population (large Centers) to smaller (Rural Community). Communities without published population classification were called "unclassified". If in the future the census publishes their population data, then these communities will have to be classified:



For example, for the Small Rural classification (10,000 < Pop < = 50,000; 200-500km from a larger city), two layers were used: point layer representing the community location and buffer zone representing the distance component.

Layer: Communities

a) Identifying Small Centers

[CSDpop2011]>=10000 AND
[CSDpop2011]<=49999 OR
[PNPop2011]>=10000 AND
[PNPop2011]<=49999 OR
[DPLpop2011]>=10000 AND
[DPLpop2011]<=49999 OR
[POPCTRpop2]>=10000 AND
[POPCTRpop2]<=49999

b) Identifying Large Centers

[MUNcategory2011] = 'Metropolis' OR
[MUNcategory2011] = 'Medium' OR
[MUNcategory2011] = 'Large' OR
[MUNcategory2011] = 'Very Large'

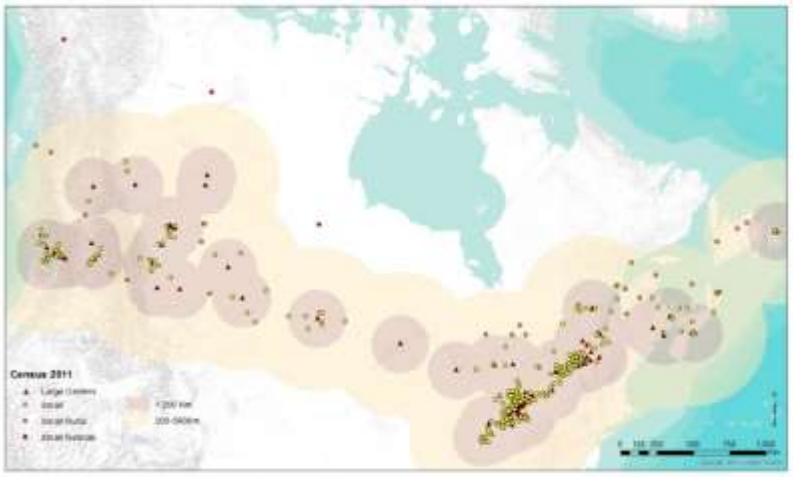


Figure 2 – Shows the Buffers used to classify the Communities

Layer: Buffer

[distance] >= 200 AND [distance] <= 500

2.4 Develop "Communities" Table

The main goal of the Communities table was to create a list with distinct communities per province and their respective attributes: census information, Lat/Long and community size classification (see item 2.3) for the Census 1996, 2001, 2006 and 2011.

2.5 Define Criteria to Integrate "Dynamic Data" with "Static Data"

The longitudinal database contained "Dynamic Data" (e.g., community size, temporal MD training experience/practice locations data) and "Static Data" (e.g., MD Student's high school/prior university locations). The model was defined to integrate both static and dynamic data using Communities table (see item 2.4). Table 3 explains how census data was considered. For example, consider a cohort where students graduated in 2006:

Table 3 – Linkage Static and Dynamic data

Medical Graduates - Class of 2006

Location of High School	Location of Prior University Attended	Location of Medical Training Experience				Location of Practice Location and Career Choice				
Admission Data		Medical Training Records				National Data				
[1]Prior 1998	[2]1998	Admission Year at Medical Program - 2002/03			Graduation Year 2006	[3]Canadian Post-M.D. Education Registry (CAPER)			[4]Canadian Medical Directory	
Census 1996	Census 1996	Med1 2002/03	Med2 2003/04	Med3 2004/05	Med4 2005/06	2Yrs Practice in 2009, 2010	2Yrs Practice in 2011, 2012,2013 ,2014	5Yrs Practice in 2013	2013 Report	2014 Report
		Census 2001	Census 2001	Census 2001	Census 2001	Census 2006	Census 2011	Census 2011	Census 2011	Census 2011

Notes:

[1]Assumption: High School happened 8 years before being admitted to the MD program. Dates were not collected.
[2]Assumption: Prior University happened 4 years before being admitted to the MD program. Dates were not collected.
[3]Releases physician practice locations every 2years, 5years and 10 years after the end of their Residency
[4]Releases physician practice locations annually upon request. Database has weekly data updates.

2.6 Integrate/Validate/Update "Communities" Master Table

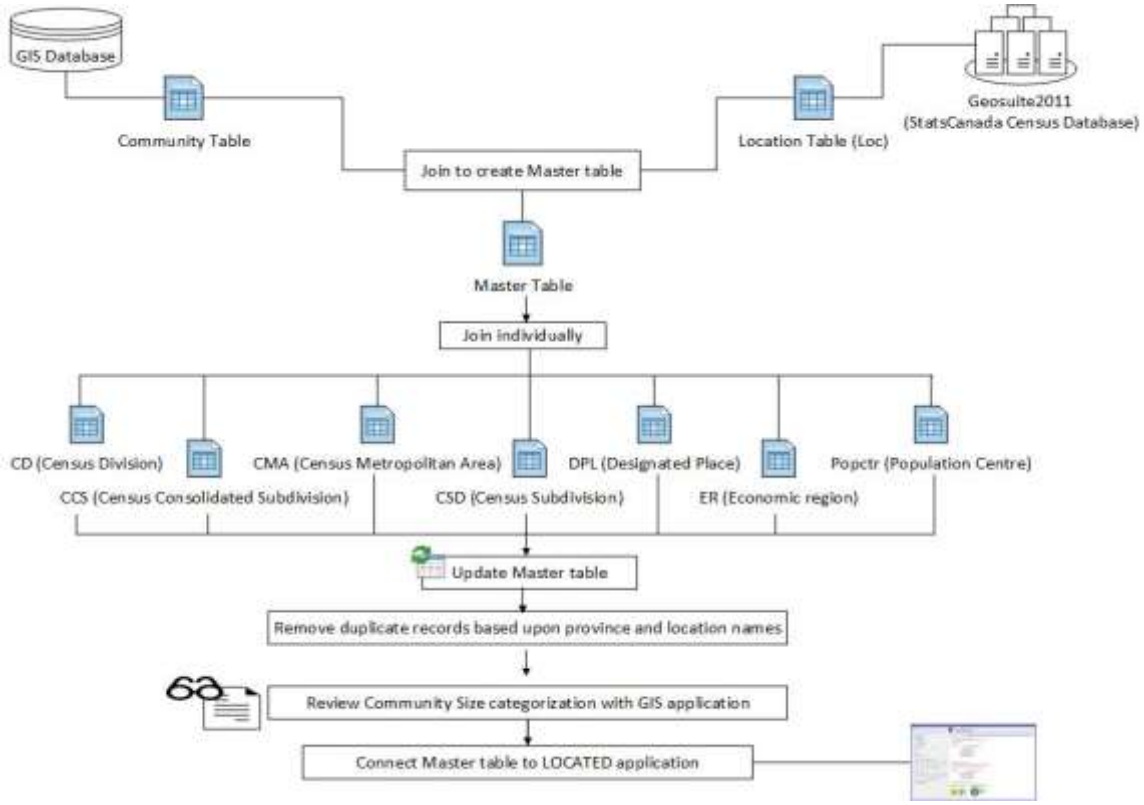


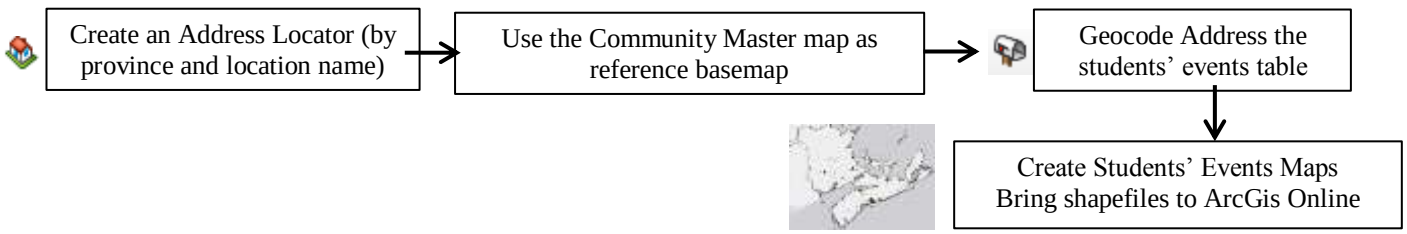
Figure 3 – Overall process to validate/update missing information on the "Community" table

Figure 3 describes the overall process to obtain the final community location information list with population categorization. The Community Table (see item 2.4) was obtained from the GIS database and integrated with the Location table (Loc) from Geosuite2011 (Census Database). Census tables such as CD, CCS, CMA, CSD, DPL, ER, Popctr were individually joined with the Community table to add/update the corresponding information that was missing from the Place Name (PN) records. The community categorization was then reviewed using the GIS application to ensure the integrity of data. Finally, list of locations with community classification was integrated with the other static-historical data (see item 2.5) to be used in the LOCATED application.

Community table from the GIS data and Loc table from Geosuite2011 were integrated on the database based upon the province and place names. The statement: *'Community Table U (Community Table ∩ Loc)'* returned all the values from Community Table along with their respective Loc table values. The table thus obtained is now the Master Table. The Master table was created using queries in PostgreSQL database management system.

2.7 Check student location using longitudinal basemap

Province names and abbreviations that were spelled differently among the data sources were tackled by creating a Converter table. This helped in accommodating the different standards used by external sources in the project thereby increasing the matching-efficiency of data around 10-15%. Once all data is integrated on Postgres, a table with the students' events is produced. Finally, the following steps were taken to represent the spatial distribution of student experiences:



3. RESULTS

This paper defines the methodology to develop a longitudinal basemap (see Figure 4). It has been created and successfully implemented on the LOCATED Project. Figure 5 displays the Large Centres to illustrate the change of the community size over time. It can be noticed that in some cases the population grows or shrinks in size.

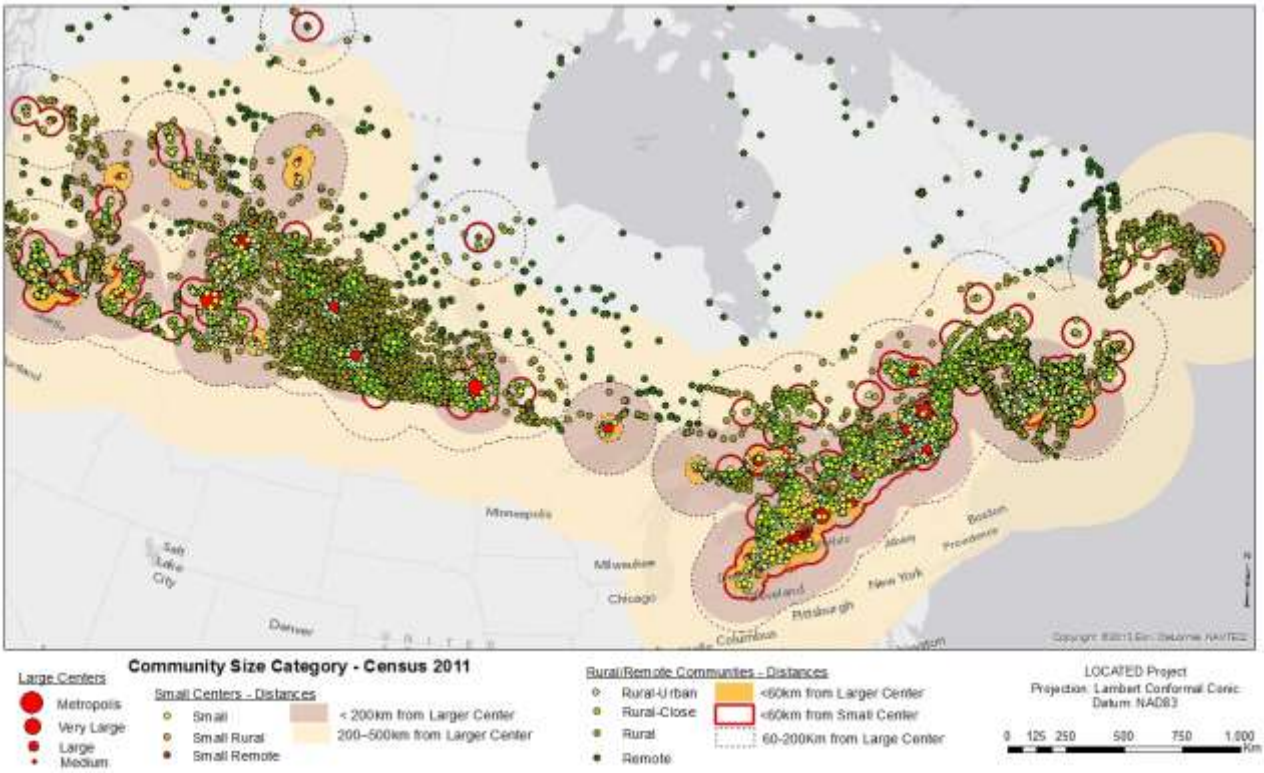


Figure 4 – Community Categorization for Census 2011

4. CONCLUSION

The implementation of longitudinal basemap is a complex task. It involves a one step-multiprocess approach. The solution to create converter spelling tables has helped to increase matching efficiency of the location/ province names (from around 80% to 95%). Once new students' data is added new unmatched location/province names are bound to occur due to misspelling or spelling variations. The converter table would then have to be updated.

So far, the methodology of the basemap development has not tested new students' event location data; it will happen post August 2014. Nevertheless, it is possible to guess that from the 6,760 locations contained on the Community Master Table, at some point new locations will have to be added; it may impact the 11 Community Size Classification (i.e., level of rurality or urbanization) too. It is a cyclic process that needs to be verified regularly. One way of avoiding dealing with the constant change in the 11 Community Size Classification, was to group them into 3 larger groups namely: **Large Centers, Small Centers** and **Rural Communities.**

Due to the small data sample collected so far: N=180 for High School location, N=272 for Prior University locations and N=181 for Practice Locations as of July 2013, aggregation of the communities' categories, were particularly important because it narrows down the chance of identifying an individual student (even though the data is de-identified). The more granular the community categories are, the more are the chances of identifying the smaller number of individuals within that group through social media.

Acknowledgment
This research is supported by the advisory committee composed by Dr. Preston Smith, Dr. Evelyn Sutton, Dr. Kathleen MacPherson, Dr. Tim Fedak, Peggy Alexiadis Brown and Gregory Power.


Census 1996
Medium: Pop: 50,000 – 100,000
Large: Pop: 100,000 – 500,000
Very Large: Pop: 500,000 - 1,000,000
Metropolis: Pop: >1,000,000


Census 2001


Census 2006
Medium: Pop: 50,000 – 100,000
Large: Pop: 100,000 – 500,000
Very Large: Pop: 500,000 - 1,000,000
Metropolis: Pop: >1,000,000


Census 2011
Medium: Pop: 50,000 – 100,000
Large: Pop: 100,000 – 500,000
Very Large: Pop: 500,000 - 1,000,000
Metropolis: Pop: >1,000,000

REFERENCE

ANDER-PECIVA, S. **Construction of longitudinal databases - for flexibility, transparency and longevity**. Centre for Population Studies / Demographic Data Base. 2005.

BARNARD, D; HU, W. The Population Health Approach: health GIS as a bridge from theory to practice. **International Journal of Health Geographics** 2005, 4:23 doi:10.1186/1476-072X-4-23.

BOULOS, M.. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. **International Journal of Health Geographics**. 2004; 3: 1.

CURTIS,A. LEITNER, M. **Geographic Information Systems and Public Health: eliminating perinatal disparity**, British Cataloguing, 2006.

DRAKE, D. **Examining the Issue of Maldistribution of Physicians through GIS: A Case Study of Retina Specialists in the United States**, 64ª Annual Meeting of the Southeastern Division-Association of. American Geographers, 2009.

ESRI, **Canada: Population Density**, 2013 Disponível em: http://maps.esri.ca/arcgis/rest/services/StatsServices/PopulationDensity/MapServer. Acesso: 15 Junho 2014.

HART, A.; MCCULLOCH,B.; HARPER, C.; GARDINER, N.; RUTHERFORD, S.; BAKER, P.; HARRIS, P.; O'SULLIVAN, D. **Report on GIS and public health spatial applications,** Queensland Health. Brisbane, 2005.

HIGGIS, G.; GOULD, M.I. Is there a role for GIS in the news"NHS". **Health and Place**, 7, p.247-259, 2001.

HIXON, A.; BUENCONSEJO-LUM, L.; RACSA, C. GIS Residency Footprinting: Analyzing the Impact of Family Medicine Graduate Medical Education. **Hawai'i Journal of Medicine & Public Health** 2012;71(4 Suppl 1):31-39.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Divulga as estimativas populacionais dos municípios em 2011**. Disponível em: http://saladeimprensa.ibge.gov.br/noticias?view=noticia&id=1&busca=1&idnoticia=1961. Acesso: 2 Junho 2014.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Metodologia do censo demográfico 2010** / IBGE. - Rio de Janeiro : IBGE, 2013. 712 p. - (Relatórios metodológicos, ISSN 0101-2843 ; v. 41).

MCCARTNEY, D.. **Country Pasture/Forage Resource Profile: Canada.** Disponível em: http://www.fao.org/ag/agp/AGPC/doc/Counprof/Canada/Canada.html. Acesso: 15 Junho 2014.

MOHAMMED,R.; LIGHT,J; MORLEY,S.; SOPER,R.; POWER,G. PAIXAO, S.; STEEVES, J. **Technical Challenges for Longitudinal Data Integration**.12th Canadian Conference on Medical Education, 2014.

PAIXAO, S.; HIPPE, J.; ROURKE, J.; STEEVES, J. **Using GIS technology for research and program evaluation at Dalhousie and Memorial: Lessons Learned**. 12th Canadian Conference on Medical Education, 2014.

PHILLIPS JR, R.;KINMAN, E.; SCHNITZER, P.;LINDBLOOM, E; EWIGMAN,B. Using Geographic Information Systems to Understand Health Care Access. **Arch Fam Med**/Vol 9:*971-97.* 2000.

PITBLADO, J.R.; PONG, R.W.;. **Geographic distribution of physicians in Canada**. Health Canada; 1999.

PONG, R.W.; PITBLADO, J.R. **Geographic Distribution of Physicians in Canada: Beyond How Many and Where**. Ottawa: Canadian Institute for Health Information; 2005.

SHIRAYAMA,Y.; PHOMPIDA, S; SHIBUYA, K. Geographic information system (GIS) maps and malaria control monitoring: intervention coverage and health outcome in distal villages of Khammouane province, Laos. **Malaria Journal** 2009, 8:217 doi:10.1186/1475-2875-8-217.

SOMAIN, R. **A população do Brasil em 2010**, 2011. Disponível em: http://confins.revues.org/7215. Acesso: 20 Junho 2014.

STATISTICS CANADA. **Population, urban and rural, by province and territory (Canada), 2011.** Disponível em: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo62a-eng.htm. Acesso: 1 Junho 2014.