

Canadian Century Research Infrastructure

Research Paper 2: Sampling Choices and Implications

Justin Ngan
Byron Moldofsky

0. Introduction.....	1
1. Spatial Statistics: Implications for Sampling Design and Choices	1
1.1 Aspatial Sampling Methods.....	1
1.1.1 Random Sampling.....	1
1.1.2 Systematic sampling	2
1.1.3 Stratified sampling.....	2
1.1.4 Cluster sampling	3
1.2 Spatial Sampling Considerations.....	3
1.2.1 Spatial Dependence.....	3
1.2.2 Effective Sample Size	5
1.2.3 Semi-variogram: Improving the Sampling Scheme	5
2. The proposed CCRI sampling approach.....	9
2.1 Sampling Design Options	9
2.1.1 The Index and Count.....	9
2.1.2 A Systematic Sample	10
2.1.3 A Clustered Sample	11
2.1.4 A Random Sample and some Empirical Evidence	12
2.1.5 Probability	15
2.1.6 Summary	18
2.2 Considerations and Alternative Approaches.....	18
2.2.1 Over-sampling	18
2.2.2 Understanding the effects of Aggregation from the 1911 sample.....	18
2.2.3 Aggregation – A Slight Re-composition of Canada	19
2.2.4 Interaction of Stratification, Clustering, and Aggregation.....	19
2.2.6 Preserving fine resolution Geographic identifiers	20
3. Methods of Spatial Analysis	21
3.0 Global and Local Statistics	21
3.1 Perspective of Spatial Analysis in Historical Microdata.....	21
3.1.1 Spatial Data and Spatial Models.....	21
3.1.2 Spatial-temporal analysis	22
3.1.3 Exploratory Data Analysis	24
3.2 Microdata and areal representation	24
3.2.1 Census Geographic Entities	25
3.2.2 Cross-Area Aggregation	25
3.2.2.1 Tobler	25
3.2.2.2 Gregory.....	26
3.3 Methods in Spatial Analysis	27

3.3.1 Global Statistics	27
3.3.1.1 Chi-square test	27
3.3.1.2 Join-Count Statistic	28
3.3.1.3 Moran's I	30
3.3.1.4 Spatial chi-square (Rogerson's R)	31
3.3.2 Local statistics	34
3.3.2.1 Local chi-square.....	34
3.3.2.2 Getis' G_i^* statistic.....	34
3.3.2.3 Local Rogerson's R.....	35
3.4 Summary	35

0. Introduction

The creation of a database of Canadian historical microdata by the Canadian Century Research Infrastructure has been established as a cross-disciplinary initiative in order to synergistically combine the knowledge and expertise from the various areas of the Social Sciences. The purpose of this paper is to present a geographical perspective for consideration by the members of the CCRI. The theme of this paper is spatial analysis as it applies to microdata. The paper is divided into three sections. The first is an examination of sampling strategies that may be considered during the assemblage of microdata. Emphasis is placed on the importance of sampling a population from historical census records, using and preserving information of spatial distributions where possible, to create a database that enables spatial analysis. From here, we move to an examination of the proposed CCRI sampling strategy and discuss some of the implications. Alternative strategies are suggested for consideration. Finally, in the third section, approaches of spatial analysis are presented that implement proven methods.

1. Spatial Statistics: Implications for Sampling Design and Choices

An aspatial discussion of sampling design generally involves covering random, systematic, clustered, and stratified sampling designs. It turns out that this is a good place to begin an introduction of the ideas that underlie spatial statistics. It is obvious that spatial statistics implies some need to consider the location attribute of observed events. A natural question of course is why? This section will provide an answer to that question. After a cursory review of aspatial sampling methods, an introduction to spatial statistics will begin with a review of the ideas presented by Cressie (1993) to establish a basis for considering space as a factor in optimizing sampling designs. Following this, some of the terms relevant to spatial statistics will be introduced. An excursion into how spatial processes translate into statistical analyses will conclude this section.

1.1 Aspatial Sampling Methods

Statistical sampling allows us to make inferences about a 'population' without necessarily surveying all cases of that population. Most often, it is too costly and inefficient to do so, other times, it may be altogether impossible, even outside of a timely and practical fashion. Sampling involves drawing a subset of events that are likely to be representative of the population under consideration. The process of sampling the population can be achieved through random, systematic, clustered, stratified, or a combination of these techniques.

1.1.1 *Random Sampling*

When there is little knowledge of a population and how characteristics may vary across that population, random sampling is a common approach to obtain a subset. In a random sample, all members of a population have an equal probability of being selected for a

sample. A sample of size n is drawn from a population of size N by randomly selecting a member from the population. Each subsequent member of the subset is drawn in the same way until n members are selected. Since each member of the population has an equal probability of being selected, the selection process assumes replacement. That is, each member that is drawn, is returned to the pool so that the probability for selecting the n^{th} member of the subset is the same as the first. If the same member is redrawn, the selection is repeated. Replacement must be incorporated but theoretically can be ignored, given a sufficiently large population such that the change in probability resulting from non-replacement is infinitesimally small.

1.1.2 Systematic sampling

If the form of the population permits, systematic sampling can be employed to select a sample and avoid the redrawing of members from the population. The method of drawing a systematic sample of size n from a population of size N begins with a single random selection from the first $[N/n]$ members of the population. The random position of that first member can be denoted as k . Each subsequent member of the subset is then chosen according to a rule that ensures complete but non-repetitive coverage of the population, such as $k+i[N/n]$, where i represents the i^{th} member of the sample. For example, if we had a population of size $N=100$ and we wanted to draw a sample of size $n=20$, then we would randomly draw from the first $[N/n]$ or $[100/20]=5$ members of the population. Assuming that the random position chosen was $k=3$, then the order in which members are drawn from the population would be 3, $3+1[100/20]$, ..., $3+n-1[5]$ or 3, 6, 13, ... 93, 98.

1.1.3 Stratified sampling

There will be times when variations exist in a population and the researcher knows in advance that grouping of the members of the population based on the variation is advantageous for analysis. For example, a population may vary by ethnicity and the researcher knows in advance that they are interested in conducting subsequent analysis for particular ethnicities. In such a case, the researcher would stratify the population by race, and then sample randomly or systematically from each stratum.

The researcher may further be aware that the strata contain largely different numbers but wish to have sufficient sample sizes for analysis. In such a case, the researcher may decide to oversample a particular strata. Oversampling changes the proportions or probability of selection for the strata. For example, in a population of $N=100$, it may be stratified into 30 members who are black and 70 who are white. A proportional stratified sample of size $n=20$ would mean selecting $70 * n/N$ from the stratum of whites and $30 * n/N$ from the stratum of blacks. This would yield 14 members who are white and 6 members who are black in the subset. This sample would be proportional because $14/70$ is equivalent to $6/30$ or 20%. Since the researcher knows that subsequent analysis would benefit from equally sized groups, they may choose to oversample black members by selecting disproportionate stratified samples. By choosing a sample that had 10 white members and 10 black members in the subset, they would change the proportions to $10/70$ for white members and $10/30$ for black members or roughly 14% against 33%.

1.1.4 Cluster sampling

As in the case of the CCRI, it may be known in advance that relationships exist in the membership of the population which may be valuable to preserve in the final sample.¹ Since a population is constructed of families in addition to individuals and that knowledge of family structures can provide insight about social and economic processes, it is worthwhile to preserve the relationship of families. A cluster sample preserving family relationships can be drawn by randomly selecting a member and including all members that are related to the selected member. Cluster sampling introduces the problem of dependence into the sample. For example, all members of a family are likely to report the same ethnicity. Stratified sampling becomes important because it can reduce the error introduced as a result of cluster sampling approaches by reducing the introduction of inadvertent dependence.

1.2 Spatial Sampling Considerations

A post-modernist critique of social science is that the claims made by researchers are too simplistic, failing to recognize the importance of time and space and recognizing the contingent nature of social construction (Johnston 2000). The fact that the CCRI is engaged in the construction of historical microdata to understand our past as well as contribute knowledge about the current and future of Canadian society is acknowledgement of the dynamic nature of society through time. Spatial statistics can contribute to addressing the contingency of place.

Each of the aforementioned sampling strategies can be modified to account for sampling a population distributed throughout space. Samples can be drawn randomly or systematically from a finite space. Alternatively, space can be partitioned to yield a sample stratified by area. The post-modernist critique can at least in part be addressed to the satisfaction of some. But beyond *satisficing*² the post-modernist admonishment, what is the empirical basis for considering the benefits of employing spatial statistics in the CCRI's endeavour? It is all too obvious that things natural and anthropogenic vary over space. Following is a review of the basis that has been established to support the consideration of space as a method to improve and perhaps make more efficient the traditional aspatial approach to statistics.

1.2.1 Spatial Dependence

¹ This might seem odd at first, but the value of aggregate data should not be overlooked. Aggregate census data can help in the development of the historical microdata.

² The term *satisficing* is used here purposely outside of its usual context to acknowledge that the post-modernist view is more complex and that research methods attempt to draw from the available information and expertise to make an attempt that may be sub-optimal but sufficient to make headway in the construction of knowledge about our past, present, and future. The term coined by Herbert Simon combines the words satisfy and suffice to describe firm behaviour. For more, see Simon, H. (1956), "Rational Choice and the Structure of the Environment", *Psychological Review*, 63.

The need for statistical sampling has already been mentioned; it allows for the ability to make predictions of some characteristic of a population from a smaller subset or sample of that population. Spatial sampling as the term suggests, is sampling from a population distributed over a given space. In order to conduct spatial sampling, the population must be referenced spatially at some scale, whether by province, by census district, subdistrict, or even street address or lot number. In the present discussion, we will set aside the question of scale to look at associated theoretical issues.

Given that the sample is substantially smaller than the population, representative geographic coverage by the sample is likely to be reduced. That reduction can be controlled in part by utilizing a stratified sample that partitions the given space. Partitioning the space to obtain a spatially representative sample implies that there may be some spatial process that leads to variations in characteristics of the population. To establish the framework for this excursion into spatial statistics, the following notation is repeated from Cressie's (1993) example.

Given a population as defined by $D \subset \mathbf{R}^2$, the purpose of spatial sampling is to select a sample such that $\{s_1, \dots, s_n\} \subset D$. That is, samples belong to a domain or population that are located on a Cartesian plane. From such a sample, it is then possible to make inferences about the population. A spatial planar process is then a real-valued stochastic process and inference on unsampled parts of the process can be predicted by $\{Z(s): s \in D\}$ where $D \subset \mathbf{R}^2$ (Ghosh and Srivastava 1999). At this point, we have only accounted for the general trend over space and the complications from spatial interaction of neighbouring observations left out.

Tobler's first law of Geography states that everything is related to everything else, and that things closer together tend to be more alike than things further apart (Tobler 1970). This introduces the notion of spatial dependence or the tendency of a characteristic or variable of interest to have similar values for locations which are closer together. Recognizing that spatial dependence is present can be useful for the purpose of optimizing a sampling scheme and for providing a basis of estimating values at unsampled locations. Following Cressie's notation, a model that assumes spatial dependence can be written as: $Z(s) = \mu(s) + \delta(s)$, $s \in D$.³ In this model, $\mu(\cdot)$ is the large-scale, deterministic, mean structure of the process (i.e. trend) and $\delta(\cdot)$ is the small-scale stochastic structure that models the spatial dependence among data.

A result of spatial dependence is that because members of samples that are close to each other tend to have similar values, it reduces the effective sample size, n (Cressie 1993). This is important when statistical tests are undertaken since the smaller effective sample size translates to a reduction in the degrees of freedom assumed. This has been demonstrated by Cressie (1993) and restated by others including Rogerson (2001), and Ghosh and Srivastava (1999).

³ Those familiar with this model will recognize that its application has been prominent in environmental applications such as the use of kriging to interpolate values within a continuous field. This real-valued stochastic portion of the prediction of the model can be adapted to various situations including point pattern data Cressie, N. A. C. (1993). Statistics for spatial data. New York, J. Wiley & Sons..

1.2.2 Effective Sample Size

The case is shown by comparing the confidence intervals for samples that are spatially independent and those that are spatially dependent. When members of a sample are spatially independent and the variance is σ^2 , a 95% confidence interval for μ is the familiar $(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$. If however, members of a sample are spatially dependent, then for two members of a sample, x_1 and x_2 , x_2 is no longer drawn from a normal distribution with mean μ and variance σ^2 . Rather, x_2 is now dependent on x_1 and can be expressed as $x_2 = \rho x_1 + \varepsilon$, ε comes from a normal distribution with mean 0 and variance $\sigma^2(1 - \rho^2)$, and ρ is a constant between 0 and 1 representing the dependence between sample members. Cressie (1993) shows that given dependence, the variance of the mean is equal to

$$\sigma_x^2 = \frac{\sigma^2}{n} \left[1 + \frac{2\rho(n-1)}{n(1-\rho)} - \frac{2\rho^2(1-\rho^{n-1})}{n(1-\rho)^2} \right]$$

The effect on the confidence interval is shown by considering an example for $n = 10$ and $\rho = 0.26$. Then

$$\sigma_x^2 = \frac{\sigma^2}{10} \times [1.608].$$

A 95% confidence interval for μ is then $(\bar{x} - 2.458\sigma/\sqrt{n}, \bar{x} + 2.458\sigma/\sqrt{n})$. The variance of the mean can be written as

$$\sigma_x^2 = \left(\frac{\sigma^2}{n}\right)[f],$$

where f is the inflation factor resulting from dependence. This can be rewritten with a substitution of

$$n' = \frac{n}{f} \text{ to get,}$$

$$\sigma_x^2 = \left(\frac{\sigma^2 f}{n}\right) = \left(\frac{\sigma^2}{n'}\right), \text{ where}$$

$$n' = \frac{n}{f}$$

represents the effective number of independent observations. For the example, $n=10$ and $\rho=0.26$, $f=1.608$ and $n' = 6.2$. The 10 dependent members is therefore equivalent to 6.2 independent members.

1.2.3 Semi-variogram: Improving the Sampling Scheme

Understanding the nature of spatial dependence that is in operation can be advantageous in optimizing a sampling approach. The semi-variogram shows the distribution of semi-variances between each sample point and every other point in a sample as a function of

the distance between the points (Figure 1). Points located near the origin result from sample points that are located close together and with similar values. A curve can be fitted to the distribution (Figure 2) to provide an understanding of the influence of spatial dependence on observed values of a variable. Figure 2 shows different patterns of spatial dependence that may arise. The traditional curve identifies a simple distance decay process of spatial dependence. Processes can also be periodic, repeating throughout space, or multi-spatial, operating at more than one spatial scale. An aspatial process, or one that might be representative of a uniformly distributed value over space of some observed variable with no general or local trends, is also shown.

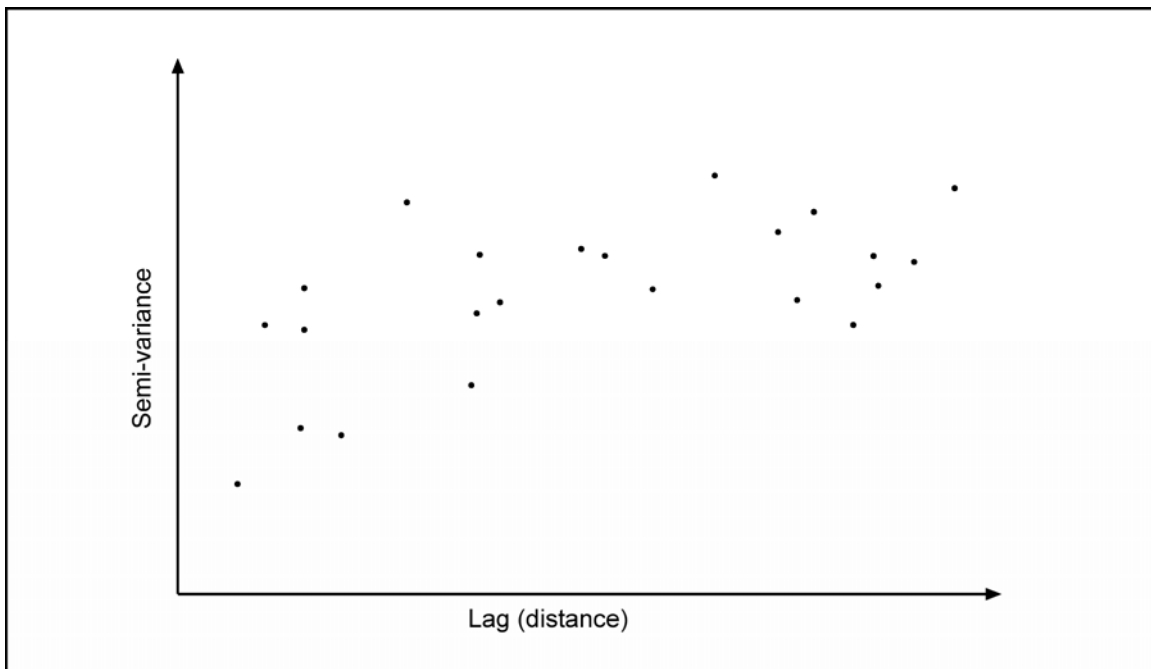


Figure 1 Semi-variance and distance in the semi-variogram

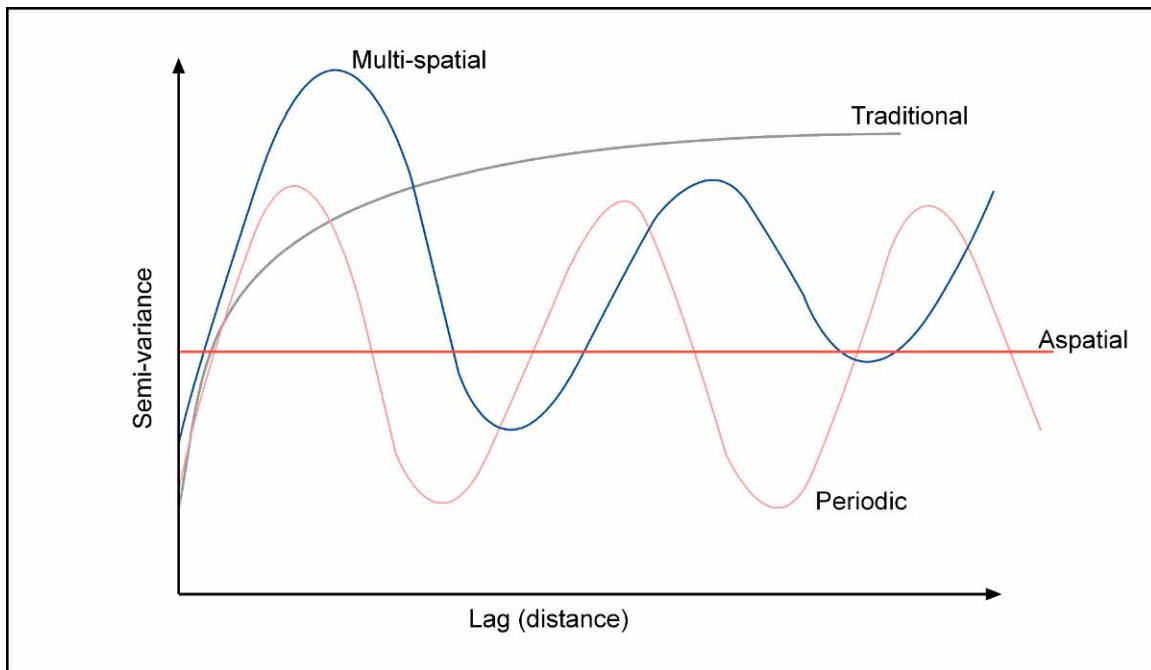


Figure 2 Hypothetical patterns of semi-variance

The semi-variogram reveals some specific details about the sample in addition to the general trend. Figure 3 shows an ideal theoretical distribution. The nugget is the point at which the graph intersects the y-axis. Sample points that are located at the same location in space might be expected to have the same observed values. This would mean that the graph should intersect at the origin. A translation along the y-axis occurs however as a result of differences from measuring the same location twice.⁴ From the perspective of census data, this can be thought of as two persons from the same household appearing in the same sample. They may have similar values such as a couple with like ages. The sill shows the maximum semi-variance in the sample. The range is the lag at which the sill occurs. The range is of particular interest since it identifies a theoretical limit under which sampling may produce redundant observations.

⁴ The notion of measuring the same location twice with different results is common in physical processes. For example, repeated measures of temperature or rainfall at a monitoring station. Minor calibration adjustments in equipment can yield different measures for an otherwise unchanged value. As applied to population studies, if we take a dwelling as the sampling 'location', two respondents may and likely will provide different responses to at least some of the questions.

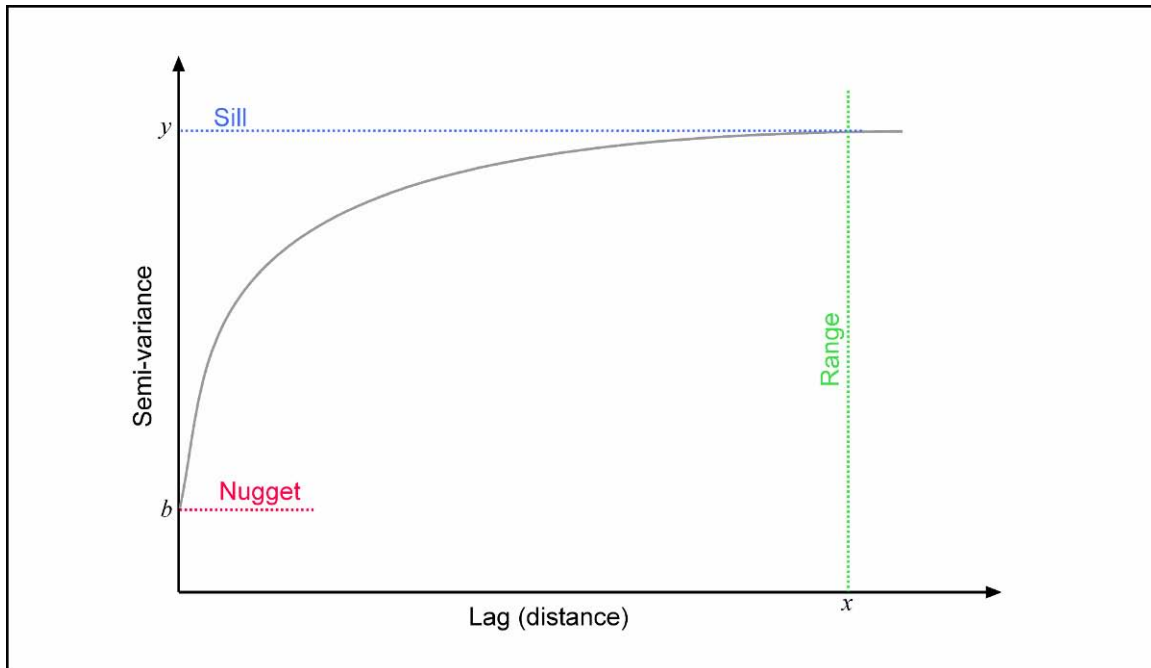


Figure 3 Structure of the semi-variogram

Given the potential for different spatial processes that might yield various spatial distributions of some census variables, and taking into account the numerous variables with which we have to work, exploring the semi-variogram can be advantageous to help in understanding the nature of the spatial patterns that are embedded in the data. It can help to identify a theoretically optimal size for sampling partitions.

While notionally, the benefits of the semi-variogram are attractive in its application to work on population data, there are some challenges. It is clear from the discussion that members of samples are geographically referenced as point locations. Practical limits prevent us from spatially referencing records as point locations.⁵ However, this does not discount its benefits or application. Rather, recognizing that the application would be constrained by the size of examined areal units, the potential to understand and exploit the knowledge of spatial dependence in the data remains. Extending the use of the semi-variogram to understand spatial dependence in population data may be an interesting pursuit with numerous payoffs.

⁵ Practical limits include the inability to geocode respondents at anything finer than census subdivisions for census in the first half of the twentieth century as well as the need to anonymize samples by aggregating to larger spatial units.

2. The proposed CCRI sampling approach

The consideration of a sampling design by the CCRI has benefited from the practical experience shared by those involved with the Canadian Families and Individual Public Use Microdata Sample projects. Despite a fair amount of expertise, the CCRI project remains far from being a simple adaptation of past approaches. The existing knowledge provides a solid foundation from which to begin but issues exist that will be specific to the landscape of 20th Century Canada. In addition, opportunities exist given the resources and teams available to enrich the final product for the broader social science research community.

The purpose of this section is to briefly review the current sampling strategy planned for 1911 by the CCRI. An assessment of the strategy will be made to identify some of the considerations of the approach, from both an aspatial and a spatial perspective. Acknowledgement is made of those who have taken the time to prepare drafts of sampling design issues for CCRI planning discussions on which this discussion is based. It is recognized that the sample produced for the first year of the time series will be constrained by unforeseeable circumstances that can be accounted for in subsequent samples. As well, this first sample can be an important basis on which future sampling strategies can be formulated as a result of astute examination of patterns, trends and other knowledge that it will provide. As such, the latter part of this section will discuss some of the future sampling questions that might be addressed based on the knowledge that will be gained in the future.

2.1 Sampling Design Options

A proposed sampling design for 1911 maintains comparability to the previous Canadian historical samples of 1871 and 1901, thus increasing the value of past and current efforts. The general design incorporates considerations of spatial stratification to obtain better geographic representation, clustering to allow examination of different social units, and oversampling to provide reasonable sizes for subsets of the sample, notably for large dwellings⁶.

2.1.1 *The Index and Count*

A preliminary step in sampling the Census manuscript will be to create a reference to the electronically captured manuscript pages. This process facilitates two important functions. First, indexing the pages is necessary to structure the data for its subsequent systematic sampling and data-entry access. Second, in creating the index, it will be possible to tabulate the number of dwellings for cross-verification and use in determining sample numbers for individual strata.

⁶ Large dwellings are those with greater than 30 persons.

2.1.2 A Systematic Sample

Given that we know little of the spatial dependency present in the population, particularly in the case where samples may be dependent on the distribution of the population among various sizes of strata such as census divisions or subdivisions, a systematic sample would seem to be a notionally effective scheme for sampling. As described in previous sections, a systematic sample would allow for a probable⁷ representative sample in which the i^{th} element selected from the population will yield similar sampling ratios when comparing the sampling ratio for the population and the sampling ratio that results for strata that may be defined for the population.

Figure 4 below shows how a systematic scheme might in theory be advantageous compared to a purely random scheme. Examination of empirical evidence follows shortly.

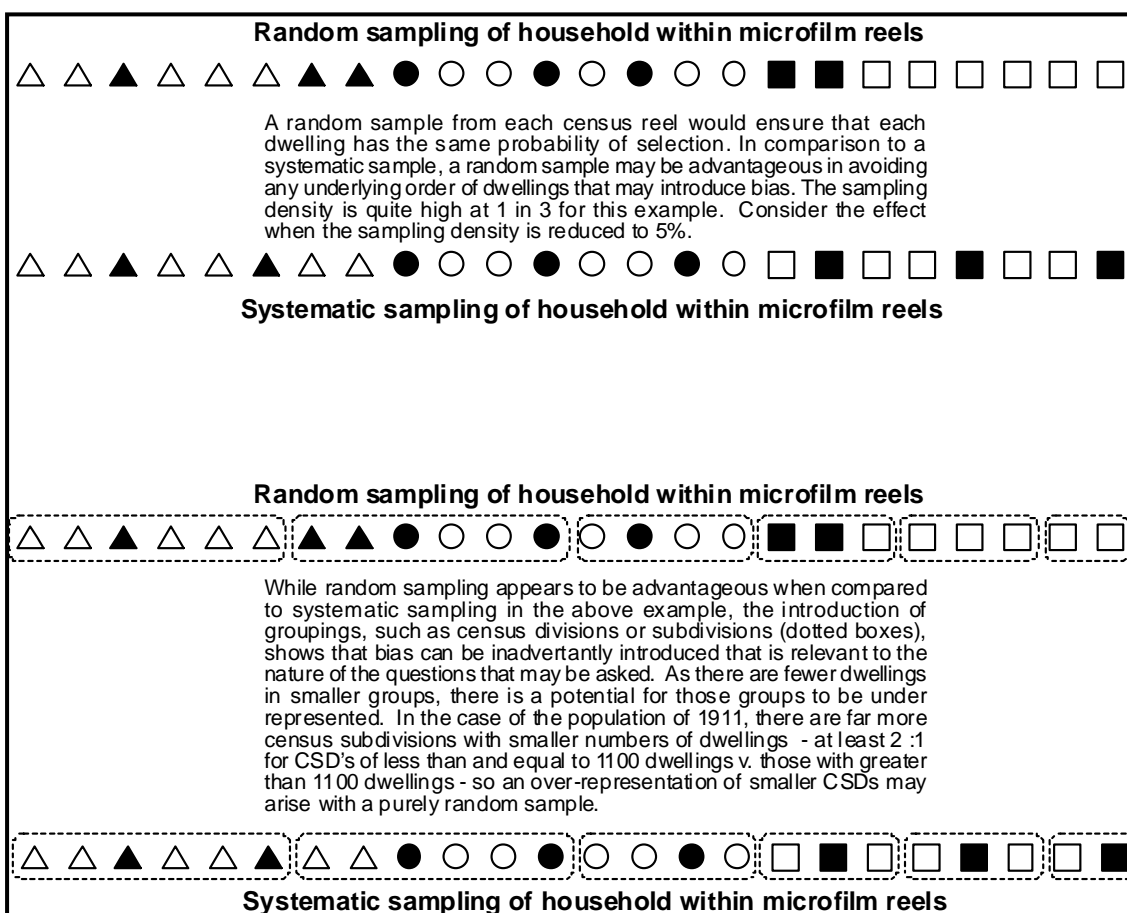


Figure 4 Comparison of Random versus Systematic Sampling With and Without Overlaid Strata

⁷ Probable though not certain because the arrangement of strata when imposed on the default order of the population list could in theory be such that particular strata will systematically be under- or over-represented. An assumption is made that the ordering of strata for some defined stratification appears in sufficiently random order to avoid such an assumed rare case.

2.1.3 A Clustered Sample

Whether systematic or random, a key design of the CCRI sample will be the ability to conduct analyses for both the population at large and for the population as constructed of families, and individuals belonging to families. This is achieved by clustering observations. The definition of dwellings are taken to be reasonably constant across sample years. So, for any dwelling selected, all members enumerated for the dwelling are included in the sample, or 'clustered'. One or more keys can be used to link all individuals in the microdata that belong to any one particular dwelling. Relationship variables can then be used to identify familial ties and non-familial relationships (e.g. borders and servants).

Error is introduced by clustering at the dwelling level because observations in the sample are no longer independent. When a particular dwelling is selected and all individuals in the dwelling are enumerated, there will be at least a few relationships between individuals that will exist. For example, a head of household of a particular race will tend to list offspring of the same race. This dependence can inadvertently increase the occurrence of particular attributes in the sample (e.g. those of a particular race).

Stratifying by geography can help to reduce cluster effects if the level of geography chosen as the stratifier captures the differences that are expressed in the clustering of samples (see Figure 5). In other words, there may be too much heterogeneity in ethnicity at one level of geographic stratification to target and overcome the clustering effects introduced at the dwelling level, such as the case of the census district level in the diagram. Meanwhile, stratifying at the census subdistrict level would allow for sampling from a subset of the population that express a different response to the variable. The benefits of clustering are difficult to refute. Particularly as we have seen changes in the family structure in response to social and economic factors introduced by the capitalist agenda of the 20th Century, the ability to conduct analyses of families and individuals as members of families is quite valuable.

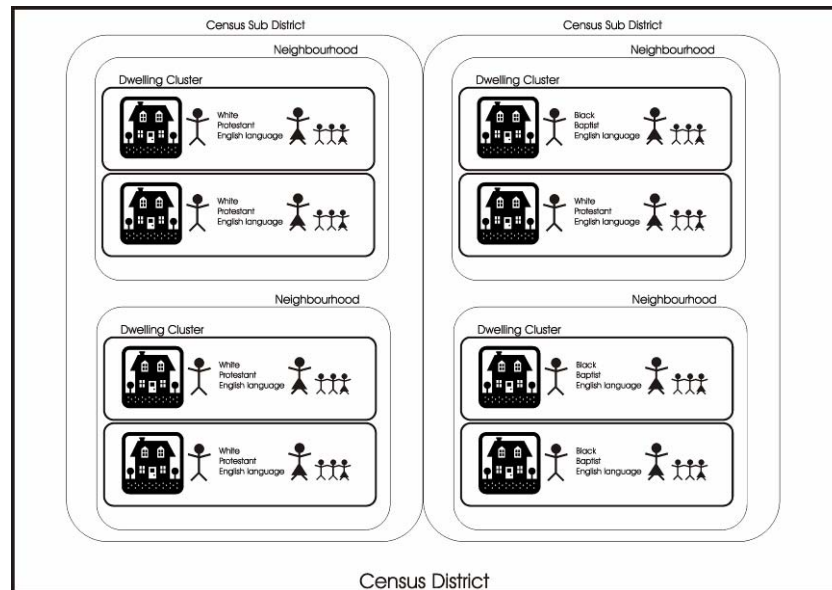


Figure 5 Overcoming cluster effects through stratification. The diagram shows how a sampling design incorporating clustering can inadvertently introduce error by over-representing particular attributes of the population. In this case, choosing a dwelling with a white head of household will result in the inclusion of the spouse and children who are also white, Protestant, and speak English. By stratifying the sample geographically, it is possible to sample populations in other areas where people of different race, religion, or language reside.

2.1.4 A Random Sample and some Empirical Evidence

Another option for sampling is to conduct a purely random sampling of the population (of dwellings). Since an index of the dwellings will have been completed as a necessary initial step to identify the population that is being sampled, that is, to define the sample space, and since we can make few assertions of the population given limited empirical evidence from research, a random sampling scheme could produce a sample of similar quality to that of a systematic sample. The combination of reasons from Figure 4 and Figure 5 for conducting a systematic and stratified sampling as opposed to a purely random one requires exploring the data for empirical evidence.

The first thing to observe perhaps is the nature of the distribution of dwellings. The distribution of dwellings can be considered by using existing data such as the aggregate census statistics. Census divisions and census subdivisions offer one way that the distribution of dwellings can be considered by using the census volumes published for the 1911 Census. Knowledge of how dwellings are distributed among these possible strata can tell us some interesting information since we know that strata containing greater numbers of dwellings often if not always are those that are representative of more densely populated urban areas. Figure 6 and Figure 7 show histograms of census divisions and census subdivisions grouped by dwelling count.

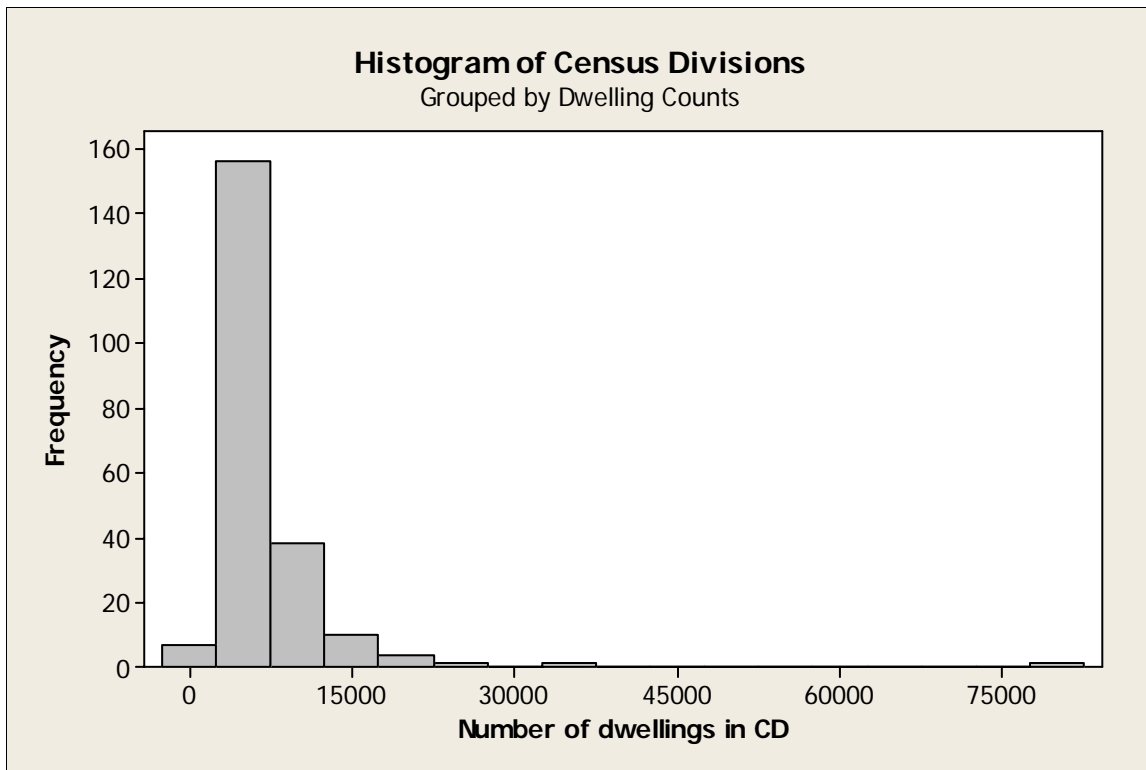


Figure 6 Histogram of Census Divisions grouped by Dwelling Counts⁸

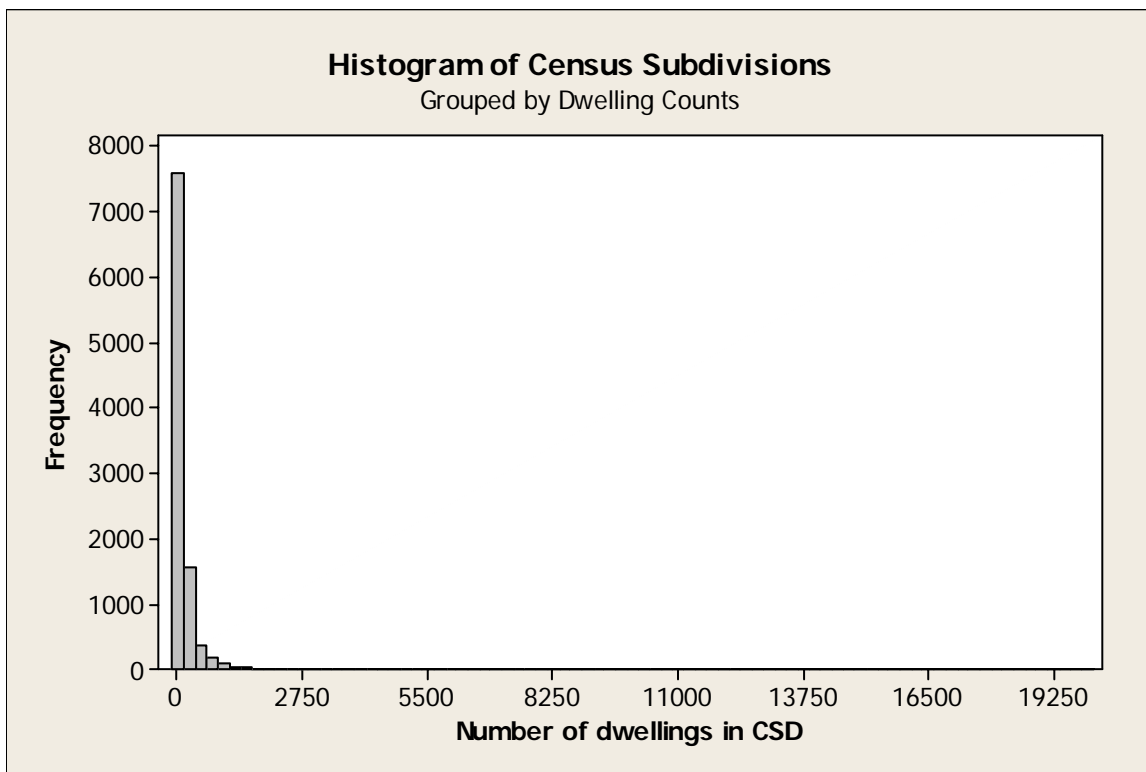


Figure 7 Histogram of Census Subdivisions grouped by Dwelling Counts⁸

In both cases, it is apparent that both distributions are not normal. The distribution of CDs has an extreme positive skew and the distribution of the CSDs is J-shaped. The graphs depict the large number of CDs and CSDs with relatively few dwellings and the dominance of the very few CDs and CSDs with substantially more dwellings⁸. Such distributions then raise a question about the relative effectiveness of a systematic versus a purely random sampling scheme. In particular, would the distribution of census divisions or census subdivisions be dependent on the selection scheme of the sample that is drawn?

To answer this question, empirical tests of the random and systematic sampling schemes were conducted to test the assumption that the size of census subdivisions is independent of the selection method. Using statistical software, random and systematic samples of 5 percent of the dwellings were drawn. The selected dwellings were tagged with the size category of the census subdivisions in which they were located. Class sizes were determined with four breaks but two more general breaks were used for the purpose of this examination. The break chosen was a dwelling count of 1100 since that was a point of natural divergence in the data distribution (see Figure 8). The results follow in Table 1 and Table 2. The tests show that regardless of the sampling scheme, the class sizes of CSDs were independent of the samples that are selected.

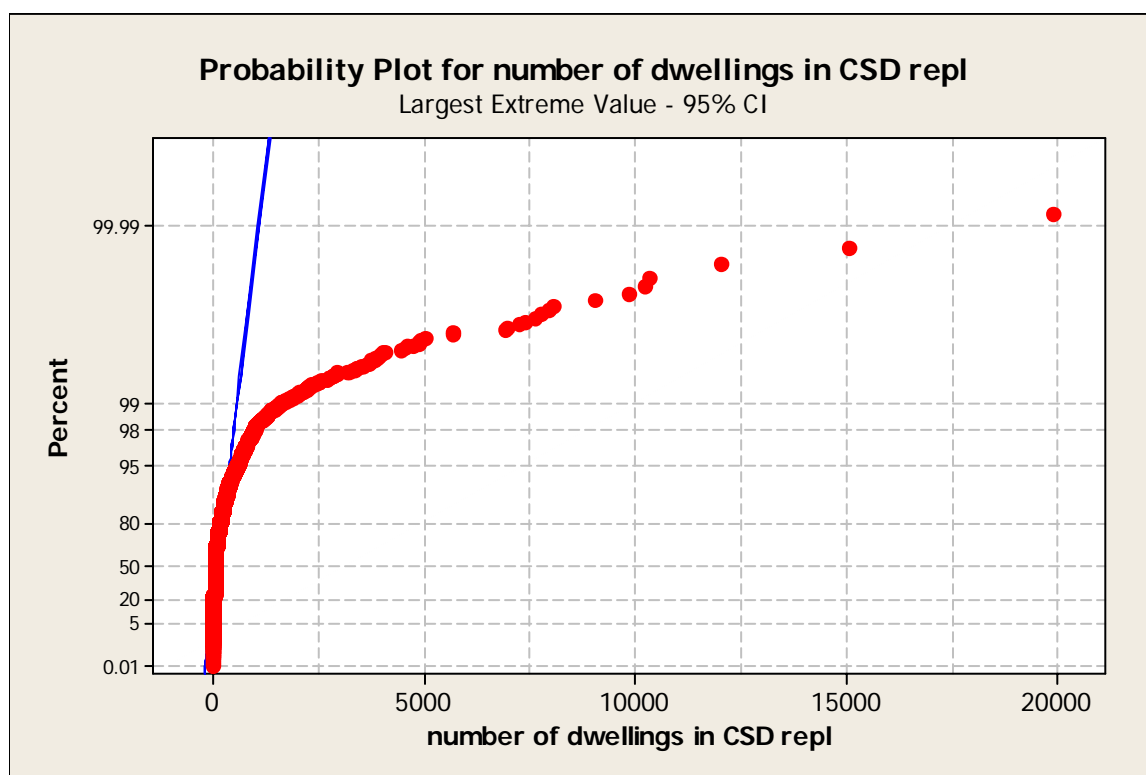


Figure 8 Class break determination

Such a result may not be intuitive. It might be expected that the small number of CSDs that contain very large numbers of dwellings might be under-represented in a random sample. The large number of selections however, offer an opportunity for the law of

⁸ Scaling of the graphs do not show accurately the bins that appear at the extreme end of the distributions along the x-axis. However, bins with at least a count of 1 exist at the far right end of each.

large numbers to come into play, and the result is that a proportion of dwellings selected for the different size groupings is similar to the proportion of for the entire population (i.e. no dependence on selection).

The purpose of considering any sampling scheme is to find the one that is most efficient. Both systematic and random samples can be drawn with similar effort. Note however that the P-value for independence drops from $P=0.99$ for a systematic scheme to $P=0.74$ for a purely random scheme. Remembering also that this empirical examination is only one of a countless number of possible samples drawn randomly, a systematic sampling scheme becomes a more appealing choice.

2.1.5 Probability

It is convenient to work with samples that have a single sampling density for the entire dataset. Such an equal probability of selection method would mean that every observation in the population (the census enumeration) has the same probability of being included. Users of the sample proceed under the assumption that every observation in the sample has the same weight. Stratified sampling introduces the possibility of samples that have different sampling densities for different strata, thus analyses that involve multiple strata require observations to be adjusted to accurately reflect their weights. Adjustments can be made using statistical software with relative ease; however it is important to highlight this point for users.

The CCRI sample will include observations that have different probabilities of selection. Specifically, observations that belong to large dwellings are over-sampled at twice the sampling density, that is, 1 in 10 as opposed to 1 in 20, thus reducing the weight of each observation from that stratum. The rationale for over-sampling the large dwellings is that even for the highest sampling density of 5% in 1911, a 1 in 20 systematic sample would only yield a single observation in a large dwelling of say 31 occupants. In addition to requirements for larger samples of the stratum for statistical analyses and analyses with the other strata in the sample, the over-sample might be more representative of the potential variations that may exist in large dwellings.

At this early stage, it seems that oversamples will be a sensible option not just for large dwellings but for particular census subdistricts that were used to enumerate large groups of individuals such as miners and railway construction crews. Even at twice the sampling density, large dwellings might not yield many more than a few observations and arguably, this may not be sufficient to capture the heterogeneity that exists nor sufficient to offer researchers viable sample sizes for analysis of these subsets of the population.

Table 1 Systematic Selection**Rows = Group; Columns = Sampling****H: hypothesis that the size class of CSDs is independent of the sample that is selected.**

Group 1: (belongs to CSD <=1100 dwellings)

Group 3: (belongs to CSD >1100 dwellings)

	Not selected	Selected	All
1	874541	46031	920572
	95.00	5.00	100.00
	66.67	66.67	66.67
	63.33	3.33	66.67
	874542	46030	920572
	-1.167	1.167	*
	0.000001556	0.000029570	*
3	437258	23013	460271
	95.00	5.00	100.00
	33.33	33.33	33.33
	31.67	1.67	33.33
	437257	23014	460271
	1.167	-1.167	*
	0.000003113	0.000059141	*
Missing	85158	4481	*
	*	*	*
	*	*	*
	*	*	*
	*	*	*
	*	*	*
	*	*	*
All	1311799	69044	1380843
	95.00	5.00	100.00
	100.00	100.00	100.00
	95.00	5.00	100.00
	1311799	69044	1380843
	*	*	*
	*	*	*

Cell Contents:
 Count
 % of Row
 % of Column
 % of Total
 Expected count
 Residual
 Contribution to Chi-square

Pearson Chi-Square = 0.000, DF = 1, P-Value = 0.992

Likelihood Ratio Chi-Square = 0.000, DF = 1, P-Value = 0.992

Fisher's exact test: **P-Value = 0.993391****Conclusion is that the size class of CSDs is independent of the elements selected.**

Table 2 Random Selection**Rows = Group; Columns = Sampling****H: hypothesis that the size class of CSDs is independent of the sample that is selected.**

Group 1: (belongs to CSD <=1100 dwellings)

Group 3: (belongs to CSD >1100 dwellings)

	Not_Selected	Selected	All
1	959757	50454	1010211
	95.01	4.99	100.00
	68.70	68.64	68.70
	65.27	3.43	68.70
	959717	50494	1010211
	39.99	-39.99	*
	0.001667	0.031677	*
3	437225	23046	460271
	94.99	5.01	100.00
	31.30	31.36	31.30
	29.73	1.57	31.30
	437265	23006	460271
	-39.99	39.99	*
	0.003658	0.069524	*
All	1396982	73500	1470482
	95.00	5.00	100.00
	100.00	100.00	100.00
	95.00	5.00	100.00
	1396982	73500	1470482
	*	*	*
	*	*	*

Cell Contents:

- Count
- % of Row
- % of Column
- % of Total
- Expected count
- Residual
- Contribution to Chi-square

Pearson Chi-Square = 0.107, DF = 1, P-Value = 0.744

Likelihood Ratio Chi-Square = 0.106, DF = 1, P-Value = 0.744

Fisher's exact test: **P-Value = 0.744096****Conclusion is that size class of CSDs is independent of the elements selected.**

2.1.6 Summary

The 1911 sample will be the first to be completed by the CCRI. While every census year will have its set of idiosyncrasies, the 1911 sample will pave the way for subsequent samples in certain respects, so keeping options open for adjustments in subsequent sampling designs, and the option to retrofit existing samples is a consideration. This suggests that variables like geographic identifiers which have implications for sampling design and subsequent analyses should be kept at as fine resolutions as possible.

It can be seen that at least in so much as it reduces the effectiveness of the sampling scheme, a purely random sample is less desirable than a systematic one. The extent to which a random sampling scheme's reduced effectiveness contributes to increasing bias in the sample depends on the strength of the association between census variables and the various strata that can be defined and overlaid to group the population. Once the 1911 microdata sample has been compiled, it will be possible to perform some empirical analysis of the associations for geographic partitions that may become common units for research and analysis.

2.2 Considerations and Alternative Approaches

As it has just been stated, variables like geographic identifiers which have implications for sampling design and subsequent analyses should be kept at as fine resolutions as possible. The following considerations and alternative approaches provide some of the justification for this statement. In section 3 of this paper, tools and techniques of spatial analyses, even more evidence will arise. For the remainder of this section, some future prospects are offered.

2.2.1 Over-sampling

Unlike the United States, there is no apparent or compelling historical basis on which to suggest over-sampling of sub-strata such as those based on race. Since the proposed sample does involve over-sampling of large dwellings, the necessary database fields to facilitate over-samples will be included, so subsequent over-samples can be integrated should the need arise.

Over-sampling should not be limited to social and economic criteria. As suggested in the first part of this paper, spatial dependence in observations can reduce effective sample sizes. The 1911 sample will provide insight into the spatial variations and processes in operation. This knowledge may highlight opportunities for over-sampling or under-sampling particular areas as part of subsequent strategies, given sufficient evidence of benefit for introducing this added complexity to the sample.

2.2.2 Understanding the effects of Aggregation from the 1911 sample

The need to anonymize the sample will require aggregation. Consideration will need to be made jointly between the principle investigators of the CCRI and those at Statistics

Canada regarding the population threshold requirements to ensure anonymity. For consideration, use of a 100,000 minimum population threshold, as in the case of the American Public Use Microdata Areas, would mean that most provinces can have several aggregated reporting units. Ontario and Quebec will have substantially more. The northern territories and parts of the Maritimes will need to be included with other aggregations since their numbers are not sufficient to form regions of their own.

We are lucky that the spatial arrangement of provinces in the country follows a west to east pattern. This coincides with the general direction of what is understood to be a global trend of westward settlement and growth. The process of aggregation should recognize this broad trend in order to capture spatial and temporal changes in the data. Where need arises, such as in Ontario and Quebec, the aggregation and zonation of population clusters should be carried out with recognition of local spatial patterns and trends that may exhibit less directionality. With the 1911 sample in place, investigation of aggregation effects in terms of zonation and scale effects should be explored so that implications of choices can be documented for researchers.

2.2.3 Aggregation – A Slight Re-composition of Canada

A primary purpose of the microdata is to gain knowledge of our past, particularly knowledge of the change from a country whose settlements were predominantly rural to one where that is mainly urban. In so doing, it may be helpful to perceive Canada beyond its provincial or common regional divisions. Such reified entities as provinces, census districts, census sub-districts, and others, may in fact limit a freer exploration of the processes that structured society. If that is not sufficiently convincing, the fact that the northern territories and some of the Eastern provinces likely do not have sufficient populations to create their own aggregations means that there will be some slight but necessary re-composition of Canada.

The suggestion made here is to explore from the information contained in the 1911 sample, whether re-composing Canada with areal units that are not bound too tightly to paradigms of political and in particular census enumeration boundaries will yield better zones from a statistical and general basis of knowledge extraction. Such aggregations may be constructed to offer more stable boundaries, thus over-coming the ephemeral nature of political and census boundaries and facilitate longitudinal comparisons.

2.2.4 Interaction of Stratification, Clustering, and Aggregation

The knowledge to be gained from the microdata can begin very soon after the creation of the first sample. The completed sample can help us to understand how stratification, clustering, and aggregation interact and what the implications are for future samples and analyses that use the data. As mentioned earlier, exploration can be made to examine the degree to which current levels of stratification help in reducing the negative effects of clustering. Where large dwellings exist, the nature of the population densities and variation within the population at the level of the stratum, may provide a sound basis on which to make decisions of over- and under-sampling. Aggregation effects will come to light as samples are anonymized. Understanding the effects of aggregation in this

particular process will provide researchers with better understanding of the data and caveats to be remembered while working with the data.

2.2.6 Preserving fine resolution Geographic identifiers

At this stage, it is too early to make decisions about the level of geographic identifiers to include. The finest possible and practical level of geographic identifier should be preserved. The retention of such information is very necessary if options for future strategies are to remain viable. For example, if a single geographic identifier is used and generalized to coarse census geographic units, the potential for re-composing the sampling frame with comparable zones for longitudinal studies will be limited.

As section three will show, some techniques in spatial analysis can take advantage of fine spatial partitions, and yield interesting and worthwhile results. In addition, this next section will also suggest a possible alternative for dealing with the issue of anonymity while offering researchers greater freedom to explore the data using recent innovations in spatial analysis tools. The key to unlocking the potential of such techniques, once again, lies in the availability of fine geographic identifiers.

3. Methods of Spatial Analysis

3.0 Global and Local Statistics

The detection of patterns in spatial analysis can occur during a search to find non-random patterns in the spatial arrangement of variable observations or as a confirmatory approach to locating specific pre-supposed concentrations, perhaps even around *a priori* identified locations. Spatial analysis procedures can be grouped using this distinction in investigative approaches. Global statistics are available to confirm the presence of spatial arrangements that differ significantly from random arrangements. Local statistics focus on confirming the presence of clusters. This chapter presents some of the methods and tools of analysis available for spatial analysis. Some basic concepts common to the various techniques are presented as perspectives of spatial analysis in historical microdata. The remaining parts of the chapter introduce ideas including: areal units for analysis, exploratory data analysis, spatial-temporal analysis, and measures in spatial analysis.

3.1 Perspective of Spatial Analysis in Historical Microdata

The availability of historical microdata will offer historians, sociologists, political scientists, and geographers the opportunity to undertake research on Canada's past with greater detail to gain knowledge and perhaps correct former misconceptions. The research will likely be diverse in nature, but common to all will be the demonstration of the regional trends or similarities and differences across the nation. Spatial analysis will yield understanding about the geographic patterns and processes of Canadian society, placing emphasis on the differences and similarities as they occurred across the country during the 20th Century. Some background is provided here as a base from which to introduce specific techniques and methods of spatial analyses.

3.1.1 Spatial Data and Spatial Models

Spatial data differs from other data in that members of the collection can be geographically referenced from one of its attributes. Conceptually speaking, the quality of geographic attributes is controlled by the potential of the attribute to distinguish the location of one record from another. Thus, an attribute, or field of a record that provides a very specific location reference is said to have a fine geographic reference, say the address on a street. Conversely, a broad reference to location, such as province, is said to be a coarse geographic reference.

Applications of spatial analysis in human geography very often entail a somewhat less fine level of geography in order to meet the need for anonymity. In the construction of historical microdata, the quality of geographic identification is also limited by the availability of a spatial framework which can geocode the location information provided. The spatial framework allows the researcher to place into a spatial context, that is, geocode, the information captured by the geographic attribute.

Data common in the study of human geography is lattice data. Such data is perceived as belonging to tessellations of the study area. This is as opposed to geostatistical or point data (Figure 9). The nature of lattice data has profound impacts on both the types of spatial analyses that can be undertaken and the interpretation of their results. Two prominent issues involve the modifiable areal unit problem and ecological fallacy.

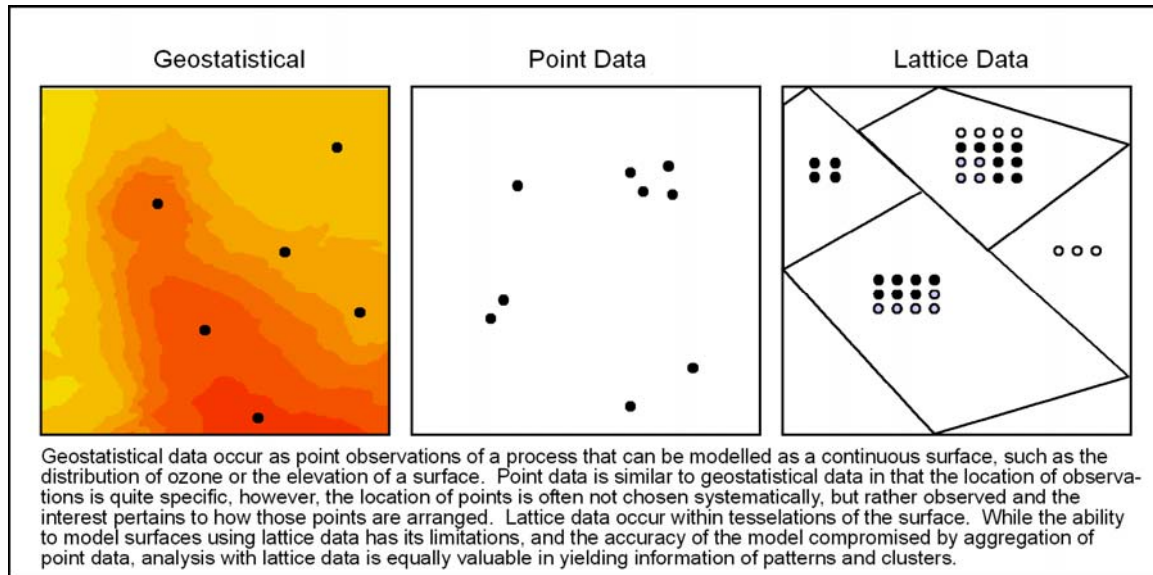


Figure 9 Types of spatial data

3.1.2 Spatial-temporal analysis

The issue of statistical dependence can arise for all analyses whether one is working with aspatial data of one time period, aspatial data of multiple time periods, spatial data observed at one time, or spatial data over multiple periods. Setting aside the problem of dependence between multiple variables that is a concern when dealing with aspatial data at one time period, this section will present the increased complexity of working with spatial data over multiple time periods.

An explanation of the difference between temporal analysis, spatial analysis, and spatial temporal analysis is described by Cliff and Ord (1981). A researcher conducting temporal analysis must be aware of the dependence between observations that stretches across time. Thus, a one way (as far as we know) relationship can be present in which an observation of an earlier time period can influence the value of the observation in a later time period. On the other hand, in working with spatial data, a researcher needs to consider that dependence occurs in more than one direction. Dependence between observations in spatial analysis can occur simultaneously between multiple observations and in multiple directions.

The complexity of modeling spatial-temporal processes has been represented by Haggett et al (1977) and reproduced in (Figure 10). In a purely temporal autoregressive study, an observation $x_{i,t}$ is a function of $x_{i,t-k}$. In a purely spatial autoregressive study, an observation $x_{i,t}$ is a function of $x_{j,t}$. When the methods are mixed, the dependency of

understand the meaning of census variables. The work of the CCRI in rectifying variables to provide maximum cross-year comparability will make working with variables less cumbersome, and it is expected that the documentation of the dataset and supplementary resources such as period articles will increase the knowledge of researchers interested in making sense of their analysis.

3.1.3 Exploratory Data Analysis

Improvements in geographic information systems (GIS) have made working with spatial-temporal data more accessible to the researcher with even modest skills. Database programs can handle a large set of records with numerous fields and allow a researcher to query and cross-tabulate the data with ease. Geographic information systems extend the functions of a database by exploiting the geographic information contained in records.

The most obvious, and perhaps significant, aspect of geographic information systems is the expansion of exploratory data analysis for research. While traditional statistics may place different emphasis on results that have *a priori* versus *post hoc* assumptions, exploratory data analysis using geographic information systems almost implies a degree of *post hoc* assumptions should be incorporated.

The relative ease to quickly display a dataset and see spatial patterns and compare patterns between time periods allows the researcher to understand the data with which they are working. Fotheringham, Brunson, and Charlton (2000) describe the need to ‘get a feel’ for the data as an initial step to research prior to the formulation of any hypothesis. A sense of the relationships between variables, their distributions, and their spatial distributions can be gained by taking this initial step. Unwin (1996) placed particular emphasis on the visualization of data as an intrinsic part of the process towards cognitive association of concepts and theories that are relevant part of a researcher’s initial considerations.

The CCRI data will entail the necessary geographic information and spatial referencing frame to allow researchers to visualize their data early in their work. The opportunity to take advantage of this option should not be overlooked, nor its benefits understated.

3.2 Microdata and areal representation

The CCRI microdata will inherit from its source, geographic reference in terms of census geographic entities, namely census districts and census sub-districts. These will represent in most cases, a practical basis for data exploration and representation given the difficulty associated with reconstructing fine geographic reference frames. Requirements for anonymity may require subsequent aggregations of this data into broader areal units. This section will present some possible methods of areal representation and aggregation when working with historical microdata.

3.2.1 Census Geographic Entities

The work of geographers in reconstructing a spatial framework that accurately portrays census boundaries for each census year will help to improve spatial analysis and visualization of the data. The spatial framework produced for the 1911 sample will include census districts and census sub-districts reconstructed through limited historical maps and substantial work to provide the boundaries in digital form. These digital boundaries will enable spatial analysis and data visualization using GIS along with advanced tools and techniques.

Beyond the reconstruction of census boundaries for each period, geographers, with the support of knowledge from the other disciplines in social science, can provide consistent spatial boundaries to facilitate studies of change across space and time to address the problem of shifting census boundaries. Experience in this area will also be a benefit when working with supplementary datasets that are not part of the CCRI's spatial framework and which may have different spatial boundaries.

3.2.2 Cross-Area Aggregation

Cross-area aggregation is a problem when attempting to compare areas with shifting boundaries over time. Certainly, one way to handle shifting boundaries is to choose larger aggregations or census units that are more stable, but this comes at the cost of reducing what is arguably an already rather coarse level of geography. In addition, shifts and changes in boundaries are apparent even at the large provincial level. Methods for handling cross-area aggregation to produce boundaries not consistent with historical divisions in census units but consistent throughout time can offer an approach that provides a manageable way to examine data across the different census periods without significant loss of regional and local detail. Methods have been documented by Dempster et al(1977), Flowerdew(1988), Flowerdew and Green(1991), Gregory(2000), and Tobler(1979; 1991). A more specific introduction to some of this work is presented in the remainder of this section.

3.2.2.1 Tobler

Tobler's solution to the problem of cross area aggregation is in effect to refute that there is any problem. Instead he suggests that the problem lies in the inappropriate choice of statistic which should yield invariant or predictable results across different aggregations (Tobler 1991). Such statistics would thus yield frame independent spatial analysis.

Tobler identifies two parts to the modifiable areal unit problem which he calls the partitioning problem and true aggregation. The partitioning problem is related to the possible alternative divisions of a space into N units. Tobler's suggestion of considering to what extent the associations between the attributes differ for the two partitionings may be unclear.

The partitioning problem also includes the scenario in which two datasets have different spatial partitions and, the solution is to find a larger aggregation in which the boundaries of the two coincide. This he cites as being common with political and hierarchical boundaries. Rather than aggregation, the recommended approach is to convert the areal

data from the spatial frame of one dataset to the other through pycnophylactic interpolation (Tobler 1979). The process can produce a smooth continuous surface from areas of irregular polygons which may be more reasonable given the nature of spatial dependence and the existence of people and processes that in the the real world, are not necessarily bound by administrative boundaries. This process can be useful for those working with the CCRI data because it is recognized that the processes studied are often not spatially discontinuous. However, as Gregory (2000) notes, the 'pcynophylactic criterion' can be limiting as it requires the target units of areal interpolation to be contained within the source region.

The second part of the modifiable areal unit problem, the true aggregation problem, involves the grouping of areal units to form new larger units. Tobler questions why one might want to do this given the reduction in resolution (Tobler 1991). The CCRI faces a practical example of this scenario. For the more publicly accessible versions of the data, it may be necessary to aggregate some regions to preserve anonymity. In such situations, Tobler's point is that aggregation is not a problem as long as results are predictable; predictability can be achieved by including information about the areal units in the modeling process (Tobler 1991).

3.2.2.2 Gregory

Gregory's work with the Great Britain Historical Database and Great Britain Historical GIS is particularly relevant for consideration here because he has addressed some of the practical problems which have garnered less research attention. One of these is the examination of what techniques to employ in areal interpolation of data and the accompanying considerations.

The Great Britain Historical GIS has a substantive collection of digital boundaries representing the various hierarchies of their political administrative regions. One challenge that has faced researchers there is that the coarse district level has a great level of diversity in attributes, but they do not lend themselves to longitudinal studies because of the frequent boundary changes (Gregory 2000). Gregory's examination of accuracy in areal interpolation is based on the objective and assumption that data should be compared at the finest level of geography that is possible as cited from Goodchild and Gopal (1989).

Gregory designed and examined the effects of a dasymetric technique, using known information about population at the target unit level to control the weighting and interpolation process. After comparing these results to those of the basic areal weighting or those employing the EM algorithm, he found the dasymetric technique to produce the most accurate results (Gregory 2000). His findings also noted the importance of choosing target zones. Specifically, the size of zones is an important consideration since larger zones will result in loss of detail while smaller zones will lead to increased error in the interpolation process.

In the context of the CCRI work, Gregory's findings are significant because user-defined target zones are likely to lead to lower error. This can be a benefit in an undertaking to create longitudinally uniform areal units, giving greater freedom to the design involved in

such an undertaking. An anticipated challenge would be to define temporally stable areas for which control data such as population is available, otherwise, the original problem is iterated at yet another level.

3.3 Methods in Spatial Analysis⁹

This final section will present some of the specific methods of spatial analysis for areal data. The methods are grouped into global and local statistics. As described earlier, global statistics are appropriate for identifying trends that appear when looking at the entire study area, while local statistics are appropriate for identifying clustering at relatively small locations in the study area.

3.3.1 Global Statistics

3.3.1.1 Chi-square test

Suppose that we have conducted a regression of family size on family income. We would like to see if the distribution of the residuals tends to differ by census sub-district beyond what is expected by the occurrence of observations. Knowledge of the spatial autocorrelation can allow us to partial out its effects in other tests and in addition, it may lead us to conduct further investigation of the familial characteristics at the sub-district level. The chi-square test is useful if not only for its simplicity. However, while the test can indicate deviation from uniformity, it reveals limited information regarding the nature of the deviation and what patterns there may be.

A chi-square test can be used to test the null hypothesis that there is no pattern to the allocation of the residuals. The chi-square statistic is

$$X^2 = \sum_{i=1}^n \frac{(O - E)^2}{E}$$

O is the observed frequency and E is the expected frequency. Expected values are equal to row and column totals divided by the overall total. An example of this test is provided in the following tables.

Hypothetical Residuals

	Census Sub-district			Total
	1	2	3	
+	10	6	7	23
-	6	15	7	28
Total	16	21	14	51

Adapted from: (Rogerson 2001)

Observed and expected frequencies of residuals

⁹ Examples have been adapted from Rogerson, P. (2001). *Statistical methods for geography*. London ; Thousand Oaks, Calif., SAGE Publications. Where appropriate, the circumstances of the example have been adjusted to place the example in a context closer to the work that may be carried out with historical microdata.

	Census Sub-district			Total
	1	2	3	
+	10 (23 x 16 / 51 = 7.22)	6 (9.47)	7 (6.31)	23
-	6 (8.78)	15 (11.53)	7 (7.69)	28
Total	16	21	14	51

Adapted from: (Rogerson 2001)

$$X^2 = \frac{(10 - 7.22)^2}{7.22} + \frac{(6 - 9.47)^2}{9.47} + \frac{(7 - 6.31)^2}{6.31} + \frac{(6 - 8.78)^2}{8.78} + \frac{(15 - 11.53)^2}{11.53} + \frac{(7 - 7.69)^2}{7.69} = 4.40$$

Using $\alpha=0.05$ and 2 degrees of freedom, the critical value of 5.99. Since our test value is 4.40, we fail to reject the null hypothesis that there is no pattern.

3.3.1.2 Join-Count Statistic

The join-count statistic reveals information that the chi-square does not. Continuing the example used to demonstrate the chi-square statistic, we now want to know whether there is any pattern in the unexpected distribution of residuals. The join-count statistic confirms whether clustering is significant.

Like the chi-square statistic, the join-count statistic compares expected to observed values. In this case, the comparison is of the types of joins that exist between areal units. In this example, joins between areas would be categorized as one of ++, --, or +- (where ++ indicates a join between two areas that both have positive residuals). The procedure then involves, tabulating the number of +- joins, determining the expected number of +- joins, calculating the variance of the number of +- joins, calculating a z-statistic, and checking to see if the z-statistic is outside the range set by a chosen α -level. The expected number of joins is determined from:

$$E[+-] = \frac{2JPM}{N(N-1)}$$

where J is the total number of joins, P is the number of areas with positive residuals, and M is the number with negative residuals. N is the total number of areas and so N is also equal to $P+M$.

The variance of the number of +- joins is given by:

$$V[+-] = E[+-] - E[+-]^2 + \frac{\sum_i L_i(L_i - 1)PM}{N(N-1)} + \frac{4[J_i(J_i - 1) - \sum_i L_i(L_i - 1)]P(P-1)M(M-1)}{N(N-1)(N-2)(N-3)}$$

where L_i is the number of joins from area i to other areas.

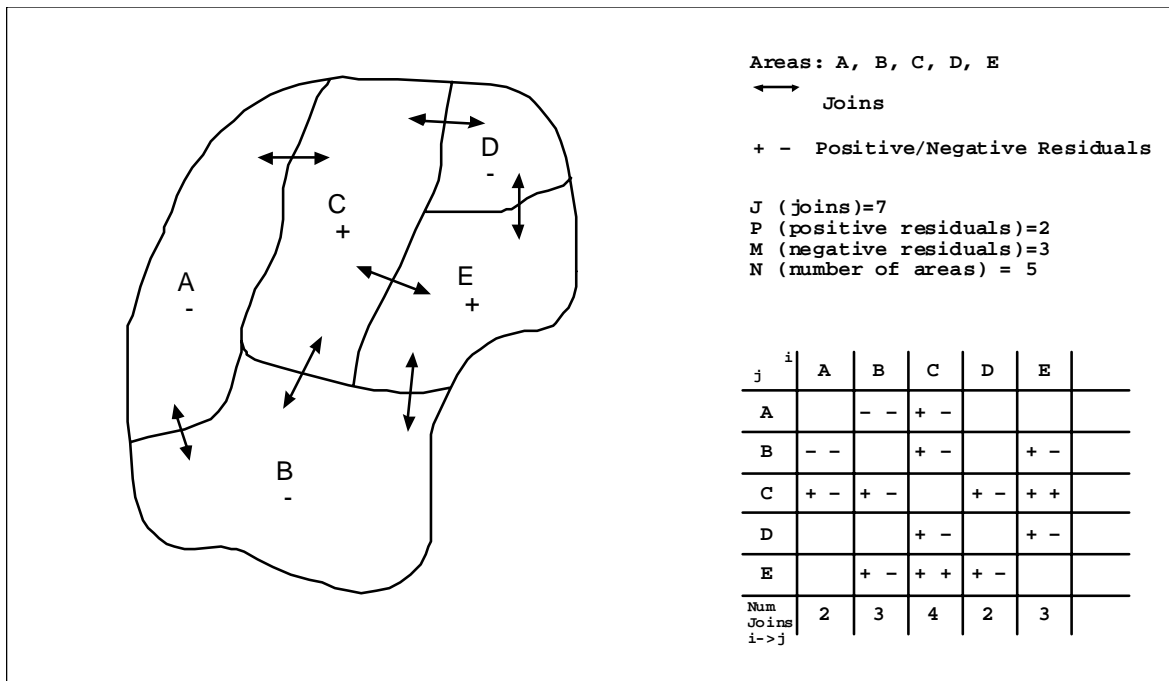


Figure 11 Example of applying the join-count statistic. Adapted from: (Rogerson 2001)

Using the equations to determine the expected number of joins $E[+ -]$ and the variance of the number of $+ -$ joins, $V[+ -]$ we get:

$$E[+ -] = \frac{2JPM}{N(N-1)} = \frac{2(7)(2)(3)}{5(5-1)} = 4.2$$

$$V[+ -] = E[+ -] - E[+ -]^2 + \frac{\sum_i L_i(L_i - 1)PM}{N(N-1)} + \frac{4[J_i(J_i - 1) - \sum_i L_i(L_i - 1)]P(P-1)M(M-1)}{N(N-1)(N-2)(N-3)}$$

$$= 4.2 - (4.2)^2 + \frac{28(2)(3)}{5(4)} + \frac{4[(7)(6) - 28](2)(1)(3)(2)}{5(4)(3)(2)} = 0.56$$

A z-statistic is calculated to test the null hypothesis that the spatial pattern is random.

$$z = \frac{(\text{Obs. "+-"} - E[+-])}{\sqrt{V[+-]}} = \frac{5 - 4.2}{\sqrt{0.56}} = 1.07$$

Since the z-statistic has a mean of zero and variance of one, a table of critical values for a normal distribution can be used. For an $\alpha=0.05$, the critical value of z is -1.96 and 1.96. Since $z=1.07$, it falls between the critical values so we fail to reject the null hypothesis that there is no spatial pattern.

3.3.1.3 Moran's I

The Moran's I statistic provides increased information about the presence of spatial autocorrelation compared with the chi-square and join-count methods. By incorporating the magnitude of observations in the calculation, unlike the join-count statistic which considers only the signs of values, Moran's I gives an indication of the strength of the spatial autocorrelation that is detected. Values for I has a range of $[-1, +1]$. '+1' indicates strong spatial autocorrelation while 0 indicates no spatial autocorrelation. Values of less than zero do not usually occur.

The formula for Moran's I is

$$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_i \sum_j w_{ij} \right) \sum_i (y_i - \bar{y})^2}$$

or using z-scores, a simpler form is

$$I = \frac{n \sum_i \sum_j w_{ij} z_i z_j}{(n-1) \sum_i \sum_j w_{ij}}$$

Moran's I offers flexibility in terms of how to incorporate and weight distance relationships. The component of the equation w is an i by j matrix ($i=j$) which allows the researcher to adjust how distance is incorporated in the calculation. The simplest form of this is a symmetric binary connectivity matrix. An example of this for the 6 region system displayed in Figure 12, is shown in Figure 13.

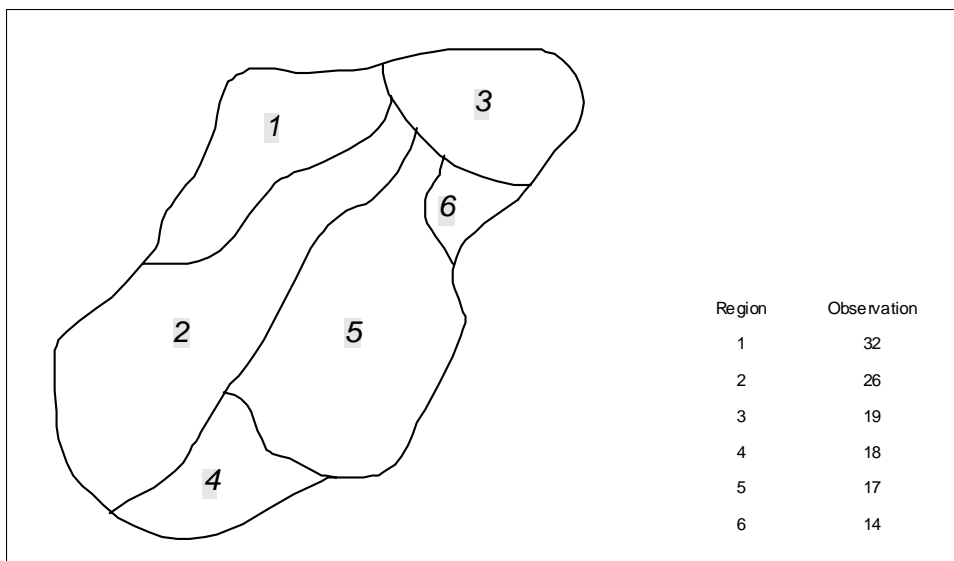


Figure 12 Hypothetical 6 Region System. Source:(Rogerson 2001)

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	1	1	1	0
3	1	1	0	0	1	1
4	0	1	0	0	1	0
5	0	1	1	1	0	1
6	0	0	1	0	1	0

Figure 13 'w' as a Symmetric binary connectivity matrix

Values along the diagonal ($i=j$) are zero to exclude comparing a region to itself. All other values are either 0 signifying no connectedness or 1 indicating adjacency. The binary feature of this scheme means that observed values are taken at face value for a comparison between pairs of areas that are connected while for those areas that are not adjacent, the 0 would exclude a comparison of the pair. Symmetry along the diagonal implies that the relationship between any pair of areas is the same whether we are going from area i to j or from area j to i , of course, this does not have to be the case and political barriers to access or gravity factors of an area can, and often do, create differential flows. From the information provided in Figure 7, Moran's I is 0.15 meaning that the spatial autocorrelation is very weak.

When more detailed information is available, the w matrix can be modified to be more sophisticated. For example, the values can represent distances between areas where the distance is calculated as the distance between area centroids. Further, a distance decay function can be incorporated so that as distance increases linearly, the weight reduces exponentially.

3.3.1.4 Spatial chi-square (Rogerson's R)

Both the chi-square and Moran's I statistics have an inherent weakness in their abilities to simultaneously assess the aspatial deviation of observations and the spatial patterns of those deviations (Rogerson 2004).

From the formula for the aspatial chi-square test, it is evident that the location of observations is not included. The first implication of this is that observations are not adjusted for spatial dependence resulting in an over-estimation of the statistic. Second, the exclusion of spatial information limits understanding of the arrangement of any poor-fit that exists in the data. Figure 14 shows two areas that can have the same chi-square value but quite different spatial arrangements for the areas with poor-fit. While the chi-square values are the same, it is apparent that the areas that have contributed to increasing the chi-square value have quite different arrangements. In the first (a), the areas with poor-fit are clustered while in the second (b), the areas with poor-fit are somewhat scattered.

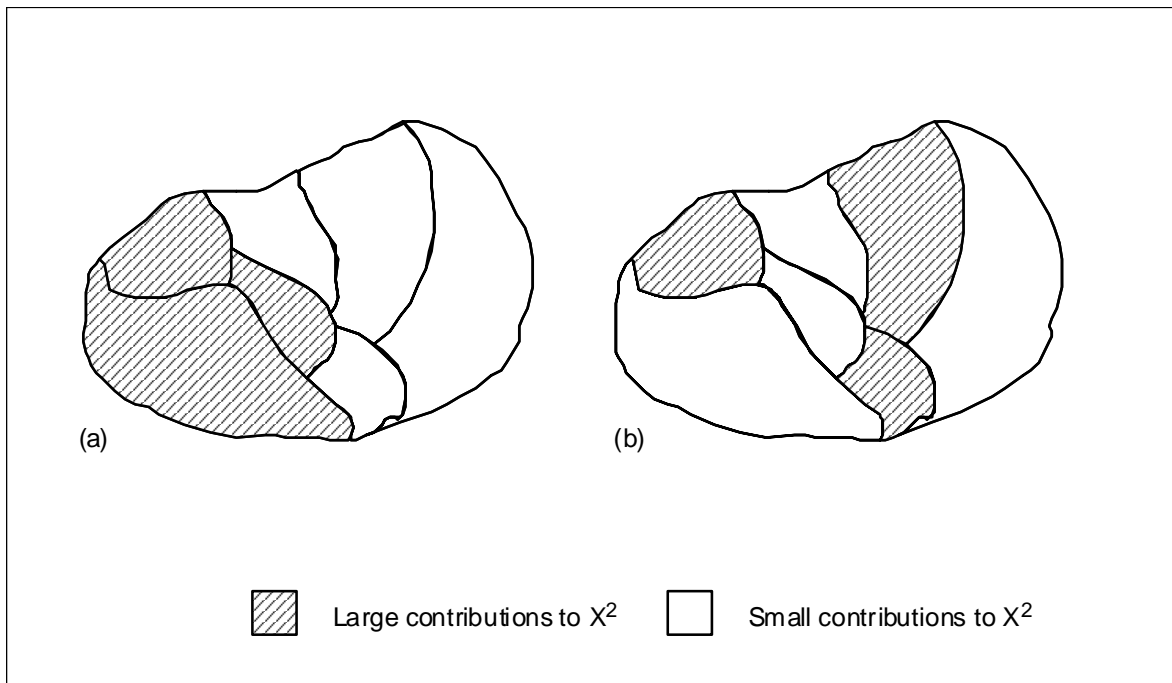


Figure 14 Identical chi-square values with different spatial patterns. Note that the X^2 for each can be the same even though the pattern of deviations is quite different, with (a) showing a clustered distribution while (b) showing one that is more distributed. Source: (Rogerson 2004)

Moran's I focuses on identifying the spatial pattern of deviations from goodness of fit, that chi-square overlooks. However, since the diagonal in the w matrix contain zeroes, a comparison of deviations within areas are not included, that is, $(y_i - \bar{y})(y_j - \bar{y})$ for all $i=j$. The importance here is to recognize that Moran's I can show the spatial patterns of deviations, but the degree of the deviations within regions is not taken into account.

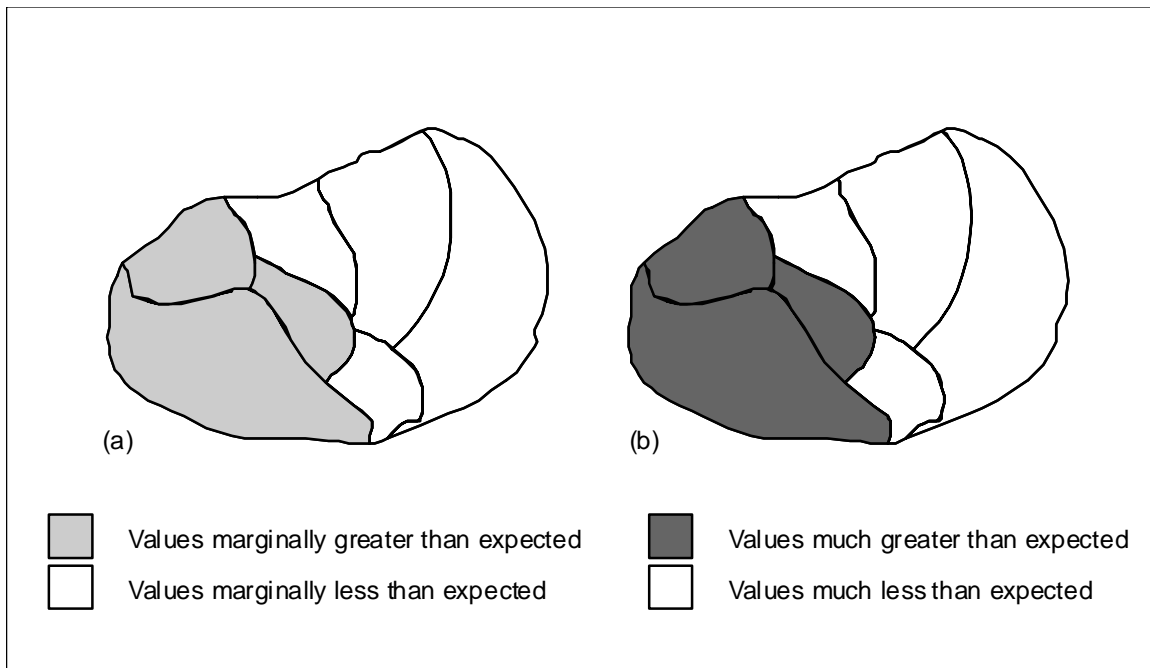


Figure 15 Identical I values with different degrees of deviations. The spatial distribution of deviations from expected values are the same in (a) and (b) as determined by Moran's I . The case on the left (a), may be not be statistically significant to reject the null hypothesis that deviations have arisen from a uniform distribution because the absolute deviations are small, while the case on the right, may be very much a rare event. Source: (Rogerson 2004)

Rogerson's R is a spatial version of the chi-square statistic that considers both the goodness of fit for the data and the spatial arrangement for areas of poor-fit. The definition of Rogerson's R is

$$R = (r - p)'W(r - p), \text{ where}$$

r is a vector of proportions for observed cases for all areas, and thus $r_i = N_i/N$;

p is the non-random vector of proportions calculated from another known variable ξ , say population, and thus $p_i = \xi_i/\xi$;

W is a matrix of weights for the statistic such that

$$w_{ij} = \frac{a_{ij}}{\sqrt{p_i p_j}}$$

The calculation of a weight matrix incorporates the importance of the connection between place i and place j . Rogerson cites Tango's use of $e^{-d_{ij}/\tau}$ which is a distance based function that incorporates a notion of limited influence around each region as determined by τ (Rogerson 2004).

The formula for calculating the spatial chi-square statistic is

$$R = \sum_i \sum_j w_{ij} (r_i - p_i)(r_j - p_j)$$

and the result can be tested against an approximate chi-square distribution with degrees of freedom as determined by

$$v = \left[\frac{(\text{Tr}(AV_p))^{1.5}}{\text{Tr}(AV_p)^3} \right]^2$$

3.3.2 Local statistics

3.3.2.1 Local chi-square

Upon discovering through the global chi-square test that there is some unexpected distribution in the variable of concern, it will likely be of interest to determine where those unexpected results occur in the data. A local version of the chi-square statistic can accomplish this.

Even when the global chi-square statistic does not show any significant deviations, one may conduct the local version of the test to determine if there are ‘hotspots’ or isolated, significant deviations that are hidden by the general trend. This might occur if there was a basis for anticipating a deviation prior to failure of the global test. Potential outliers gleaned from the data and knowledge of other processes and variable distributions are a couple of the reasons that a different outcome was expected.

The local version of the chi-square statistic is carried out by testing the significance of deviations for individual areal units. The local test presented in Rogerson(2004) is

$$z_i = \frac{N_i - Np_i}{\sqrt{Np_i(1 - p_i)}}$$

The bounds for the critical value is calculated by

$$\Phi^{-1} = \left\{ 1 - \frac{m - [m^2 - 2\alpha m(m-1)]^{1/2}}{m(m-1)} \right\}, \Phi^{-1} \left\{ 1 - \frac{\alpha}{m} \right\}$$

Φ is the cumulative distribution function for the standard normal distribution.

3.3.2.2 Getis' G_i^* statistic

Whereas the local chi-square statistic allows one to test for deviations from the expected distribution one areal unit at a time, the Getis' G_i^* statistic provides a test for the clustering of lower or higher values.

Rogerson (2001) presents Getis and Ord's (1992) statistic as:

$$G_i^* = \frac{\sum_j w_{ij}(d)x_j - W_i^* \bar{x}}{s \{ [nS_{ii}^* - W_i^{*2}] / (n-1) \}^{1/2}}$$

where s is the sample standard deviation of the x values;

$w_{ij}(d)$ is equal to 1, if region j is within distance d from region i , and 0 if not¹⁰; and

$$W_i^* = \sum_j w_{ij}(d)$$

$$S_{1i}^* = \sum_j w_{ij}^2$$

3.3.2.3 Local Rogerson's R

A local version of the spatial chi-square statistic has been presented by Rogerson (Rogerson 2004). The local version of the spatial chi-square would bring to the examination for clustering the benefit of the global version, namely the simultaneous examination of the expected distribution and its spatial pattern.

The local chi-square is defined

$$R_i = \frac{(r_i - p_i)}{\sqrt{p_i}} \sum_j \frac{a_{ij}(r_j - p_j)}{\sqrt{p_j}}$$

Given the null hypothesis of no local clustering,

$$\frac{R_i}{E[R_i]}, \text{ and } E[R_i] = \frac{a_{ij}(1 - P_i)}{N}$$

This can be tested against a chi-square distribution with one degree of freedom.

3.4 Summary

The purpose of this chapter has been to identify some introductory concepts and methods of spatial analysis. At the outset, the difference between global and local statistics or the search for trends as opposed to clusters is provided to position the nature of how geography, or space, can be a valuable component of the CCRI microdata. The nature of how that data will likely appear as observations framed within census geographic entities is then described along with some relevant concepts such as the limitations and options for working with the data. A brief description of spatial-temporal analysis is offered to establish the multi-dimensional character of historical microdata. A quick, and certainly not exhaustive, survey of some global and local measures of spatial analysis are presented as options that are open to both application and further exploration.

Research resulting from the CCRI microdata has the potential to be diverse. Researchers from the various fields of social science will no doubt have expertise and comfort with different methods, but certainly all will contribute to an understanding of what Canada and Canadians were like in the past. This modern-day looking-glass into the past is exciting because of the detail and clarity that it offers. This looking-glass can pan across

¹⁰ Note that as an alternative to a rigid definition of proximity using i , the weights can be assigned on a visual basis, perhaps a particular sub-region of interest.

the regions and help us to understand the differences that made Canada what it was, and what it is today. As each researcher crafts their own version of the looking-glass and composes an image of Canada's past, the tools of spatial analysis and presentation will be an option open to all, as a method of conveying that image to others.

References

- Cliff, A. D. and J. K. Ord (1981). Spatial processes : models & applications. London, Pion.
- Cressie, N. A. C. (1993). Statistics for spatial data. New York, J. Wiley & Sons.
- Dempster, A. P., N. M. Laird, et al. (1977). "Maximum Likelihood from Incomplete Data via the EM algorithm." Journal of the Royal Statistical Society, Series B **39**: 1-38.
- Flowerdew, R. (1988). Statistical Methods for Areal Interpolation: Predicting Count Data from a Binary Variable. Northern Regional Research Laboratory, Universities of Newcastle and Lancaster.
- Flowerdew, R. and M. Green (1991). Data integration: Statistical Methods for Transferring Data between Zonal Systems. Handling Geographic Information. I. Masser and M. Blakemore. Harlow, Longman: 38-54.
- Fotheringham, A. S., C. Brunsdon, et al. (2000). Quantitative geography : perspectives on spatial data analysis. London ; Thousand Oaks, Calif., Sage Publications.
- Getis, A. and J. K. Ord (1992). "The Analysis of spatial association by use of distance statistics." Geographical Analysis **24**: 189-206.
- Ghosh, S. and J. N. Srivastava (1999). Multivariate analysis, design of experiments, and survey sampling. New York, Marcel Dekker.
- Goodchild, M. F. and S. Gopal (1989). The Accuracy of spatial databases. London, Taylor & Francis.
- Gregory, I. (2000). An evaluation of the accuracy of the areal interpolation of data for the analysis of long-term change in England and Wales. 5th International Conference on GeoComputation, Greenwich.
- Haggett, P., A. D. Cliff, et al. (1977). Locational analysis in human geography. London, Arnold.
- Johnston, R. J. (2000). The dictionary of human geography. Oxford ; Malden, Mass., Blackwell Publishers.
- Rogerson, P. (2001). Statistical methods for geography. London ; Thousand Oaks, Calif., SAGE Publications.
- Rogerson, P. (2004). The Application of New Spatial Statistical Methods to the Detection of Geographical Patterns of Crime. Applied GIS and spatial analysis. G. Clarke and J. C. H. Stillwell. Hoboken, NJ, Wiley: xi, 406 p.
- Tobler, W. (1970). "A computer movie simulating urban growth in the Detroit Area." Economic Geography **46**: 234-240.

Tobler, W. (1979). "Smooth Pycnophylactic Interpolation for Geographical Regions." Journal of the American Statistical Association **74**(367): 519-530.

Tobler, W. (1991). Frame Independent Spatial Analysis. The Accuracy of Spatial Databases. M. Goodchild and S. Gopal. London, Taylor and Francis: 115-122.

Unwin, D. J. (1996). "GIS, spatial analysis and spatial statistics." Progress in Human Geography **20**: 540-551.