Scand J Work Environ Health 1988;14(3):168-174

<https://doi.org/10.5271/sjweh.1935>

Issue date: Jun 1988

Application of an Euclidian distance measure to the selection of reference areas in epidemiologic research concerning environmental issues.

by [Murdie RA](#), [Spitzer WO](#), [Suissa S](#)

Affiliation: Department of Geography, York University, Toronto, Ontario, Canada.

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/3393852

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Application of an Euclidian distance measure to the selection of reference areas in epidemiologic research concerning environmental issues

by Robert A Murdie, PhD,¹ Walter O Spitzer, MD, MPH,² Samy Suissa, PhD²

MURDIE RA, SPITZER WO, SUISSA S. Application of an Euclidian distance measure to the selection of reference areas in epidemiologic research concerning environmental issues. *Scand J Work Environ Health* 14 (1988) 168—174. A procedure for selecting reference areas in epidemiologic research employing census data and squared Euclidian distance is described. The procedure was adapted from cluster analysis, a multivariate statistical technique that has been applied in many disciplines. With the use of 12 census variables as the basis for evaluating sociodemographic differentiation, squared Euclidian distances were calculated between a geographically delineated index area in southwest Alberta, where residents had complained for several years about the effects of exposure to sour gas emissions, and 119 provincial census tracts in the rest of nonmetropolitan southern Alberta. The Euclidian distances can be interpreted as social distance scores with values close to zero representing a high level of sociodemographic similarity between the index area and potential reference areas. The social distance scores, in association with environmental data, suggested a clear choice for the most comparable unexposed reference area and illustrated the difficulty of finding a suitable most comparable exposed reference area. Results from the demographic component of the subsequent health survey indicated that the index area and reference area were similar in most respects. Furthermore, tests with and without statistical adjustment for confounding variables produced negligible differences on most of the important target outcome variables.

Key terms: census data, cluster analysis, epidemiologic methods, Euclidian distance.

Protocols for evaluating the health status of populations in areas affected by environmental contaminants usually call for the selection of one or more reference populations that will serve as comparison benchmarks. The problem in epidemiologic nonexperimental research is developing a procedure for identifying suitable reference populations from which a sample of participants can be obtained for health status evaluation and scientifically rigorous comparison with the affected population. Generally, this procedure involves selecting one or more reference areas that are similar demographically to the affected area, but free from environmental contaminants. In this paper we describe and evaluate a procedure for identifying reference areas that are the most similar demographically to an index area or area at risk. Methodological procedures and results from the Southwest Alberta Medical Diagnostic Review Project are used to illustrate the technique. Reference is also made to results from the Toronto Junction Triangle Study.

The Southwest Alberta Medical Diagnostic Review Project was an environmental epidemiology study mandated by the Acid Deposition Research Program of the Government of Alberta to help settle a public

health and environmental controversy (20). For nearly 30 years, residents of the communities of Twin Butte, Hill Spring, Glenwood, Mountain View, and Willow Creek (hereafter referred to as the index area) had been concerned that the emissions from sour gas plants in the vicinity were causing an excess of chronic respiratory disease, an excess of cancer, an excess of birth defects, unfavorable pregnancy outcomes, and even an excess of overall mortality. Sour gas is natural gas contaminated, mainly, with sulfur derivatives, but also with other heavy metals and trace metals. Contaminated sour gas has to be purified in a sour gas plant in order to yield "sweet gas," which can be marketed in the normal way.

Concurrent cross-sectional and cohort studies in which nearly 4 000 residents of Alberta were interviewed and examined in the summer of 1985 were designed and implemented to establish whether the residents of the index area did in fact demonstrate such excesses of adverse health outcomes. The key challenge in this nonexperimental research project was to delineate comparison or reference areas as similar as possible to the index area which had been demarcated by Alberta Environment (the department of the environment of the Province of Alberta). The unique feature of the sequence of activities was that, once the study was completed, it was possible to establish empirically whether the described method does in fact yield comparable areas.

Because of the presence in the index area of a large Mormon population with particular lifestyle charac-

¹ Department of Geography, York University, Toronto, Ontario, Canada.

² Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada.

teristics the protocol called for the identification of two reference areas, the most comparable unexposed reference area, which was as comparable as possible demographically to the index area but not exposed to sour gas emissions, and the most comparable exposed reference area, which was as comparable as possible demographically to the index area and exposed to sour gas emissions.

Selection of such areas for comparison would allow an evaluation of the effects of both demography and sour gas emission on the health problems of the index area. For example, if health problems were found to be significantly more serious in the index area and the most comparable exposed reference area than in the most comparable unexposed reference area, sour gas emission could be considered a potential cause of adverse health effects. Alternatively, if health problems were found to be significantly more serious in both the index area and the most comparable unexposed reference area than in the most comparable exposed reference area, demographic characteristics, lifestyle factors, and environmental contaminants other than sour gas could be considered potential causes.

In addition, we expected that, by oversampling the non-Mormon population in the unexposed most comparable reference area, a sufficient sample of "average non-metropolitan southern Albertans" would remain in that sample to compare with a similar group in the most comparable exposed reference area. Using the groups so assembled, we would be able to ascertain the effect of sour gas emissions on the health status of "average nonmetropolitan southern Albertans."

The Junction Triangle Study concerned an old industrial area at the junction of two major rail routes in the western part of Toronto where residents have complained for several years about the long-term health effects of exposure to industrial contaminants. The research design required the selection of a reference area in the city of Toronto that was demographically similar to the Junction Triangle but minimally exposed to industry.

Methods

Definition of the benchmark area

Statistics Canada, the national census agency, provides demographic data for a variety of spatial scales (22). However, the index area, as defined by Alberta Environment for the Southwest Alberta Medical Diagnostic Review Project, did not correspond exactly with any of the available census boundaries. For this reason, a benchmark area was defined as the most comparable aggregate of enumeration areas that corresponds with the index area. Enumeration areas are the smallest data collection units for which Statistics Canada routinely collects census data. The benchmark area had

a 1981 census population of 2 827, whereas the estimated population for the index area in 1981 was 2 228. Although the benchmark area is slightly larger in areal extent and population size, there is no reason to believe that it is not similar demographically to the index area. If the index area had been located in an urban center with a population of 50 000 or larger, Statistics Canada's geocoding system could have been used, albeit at additional cost, to match the demographic data with the exact boundaries of the index area. Unfortunately, special data tabulations of this sort are not available for rural areas such as southwest Alberta.

Definition of the comparison areas

Enumeration areas, with populations ranging from less than 100 to over 1 000, were considered too small on the average and too variable in population size for identifying comparison areas. Instead, it was decided to aggregate the enumeration area data to the level of provincial census tracts. The latter generally have 3 000 to 8 000 people, with an average of 5 000, and follow well-recognized natural or man-made boundaries. For this analysis the major advantage of provincial census tracts is their relative spatial compactness and uniform population size. The search space for suitable reference areas was restricted to the area south of an east-west line running through Edmonton but excluding the Calgary and Edmonton census metropolitan areas. Since the benchmark area was predominantly rural we excluded the two largest cities. Excluding the benchmark area, 119 provincial census tracts were included in the analysis.

Variable selection

The evaluation of demographic similarity was based on 12 variables from the 1981 census of Canada. These variables represent a wide range of demographic characteristics including age structure, family size, mobility, income, educational achievement, occupational status, and religion (table 1). In total, it was assumed that the 12 variables would provide a comprehensive demographic summary of residents in the index area and would serve to differentiate this area from other nonmetropolitan areas in southern Alberta. As a control for population size differences between census tracts, all variables except average household income were expressed as percentages. Because of the relatively large Mormon population in the index area, religion was deemed to be a particularly important variable for comparison with other parts of Alberta. Unfortunately, Mormon religion was not available by enumeration area for 1981 and therefore was not included in the formal analysis. Mormon religion was available, however, by provincial census tract for 1971, and this information was used to supplement the main analysis.

Table 1. Variables included in the analysis — 1981 census data.

Shortened name	Full description
Pop 0 to 4	Percentage of the population 0 to 4 years of age
Pop 5 to 14	Percentage of the population 5 to 14 years of age
Pop ≥65	Percentage of the population 65 years of age and over
6+ families	Percentage of the families with 6 or more persons
Mean income	Average household income
Low educ	Percentage of the population 15 years of age and over not attending school with primary education only
Some univ	Percentage of the population 15 years of age and over not attending school with some postsecondary education
Farmers	Percentage of the total labor force engaged in farming
Primary	Percentage of the total labor force engaged in primary occupations (other than farming)
United Church	Percentage of the population United Church religion
Catholic	Percentage of the population Catholic religion
Occup ≤2 years	Percentage of the private households occupying present house for two years or less

Measurement of demographic similarity

The method for measuring demographic similarity between areas was adapted from cluster analysis. Cluster analysis is a multivariate statistical technique that has been used by disciplines in both the physical and social sciences. Much of the early work was in biology (19) and psychology (4), but there are examples in fields such as geography (2) and marketing (11). Romesburg (16) has provided a recent summary of the technique and various applications. Although there is a large number of classification procedures available, the most commonly used is the stepwise or hierarchical grouping technique developed by Ward (24) and operationalized for the computer by Veldman (23). The basic objective of the cluster analysis is to classify objects (eg, skulls, tree leaves, individuals, cities, areas within cities, areas within states or provinces) on the basis of a set of relevant characteristics (eg, length, width, shape, size, personality traits, economic structure, age of population, ethnicity, occupational status). Examples of the use of the procedure in work related to this study include classifications of Canadian and American cities (10, 15), socioeconomic regionalization of census subdivisions in Ontario and Quebec (3), and classifications of census tracts in Calgary (5) and Edmonton (6).

The indicator of similarity among two or more communities or populations is expressed as a social distance score where the possible values start with the lowest score of zero, denoting perfect similarity, to very high values. The values depend on the number of vari-

ables employed in the analysis and the unit of measurement of these variables. The method used by Ward (24) for computing distances in the clustering procedure is based on the Pythagorean theorem (the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides). More generally, the squared Euclidian distance between any two observations and any number of characteristics can be determined according to the following algorithm (13):

$$D_{ij}^2 = \sum_{k=1}^n (X_{ik} - X_{jk})^2,$$

where X_{ik} and X_{jk} are the values for observations i and j on variable k , n is the number of characteristics or variables, and D_{ij}^2 is the squared distance separating observations i and j .

The 12 variables listed in table 1 were standardized with the use of Z scores in order to give each characteristic the same importance in the analysis. Squared distance (D^2) values were then computed between the benchmark area and the other 119 provincial census tracts in Alberta south of Edmonton. In effect, these scores are measures of social distance between the benchmark area and provincial census tracts. The smaller the value, the more similar the census tract is to the benchmark area. In this study, the distance measure was obtained through a modification of Veldman's HGROU computer program (23). Recently, however, it has become possible to obtain this measure with one of three widely available statistical software packages, ie, BMDP (8), SAS (17), and SPSS^x (14). The 28 provincial census tracts with social distance scores of less than 15 were mapped and rank ordered from the lowest to the highest. The social distance score is a descriptive measure for which there is no test of statistical significance. Given the nature of the data in this study, it was judged that scores greater than 15 would be difficult to handle even with statistical adjustment.

Location of sour gas plants

The location of sour gas plants in Alberta was derived from a number of sources, including Alberta Environment (1), the Energy Resources Conservation Board (9), and Delta Projects Limited (7). No attempt was made to categorize the plants by characteristics such as size, emissions, date of construction, and emission controls. An attempt was made, however, when we refined the research for the most comparable exposed reference area, to take into account the predicted areas of deposition surrounding each plant on the basis of prevailing wind direction and height of stack.

Results

Demographic comparisons for the benchmark area and the average and standard deviation values for all

provincial census tracts included in the analysis are shown in table 2. Compared with the average tract, the benchmark area had a considerably younger population, a much larger family size, a higher proportion of persons with university training, a much higher proportion of agricultural workers, and a less mobile population. It enjoyed about the same income as the average tract and had an average proportion of persons with low educational achievement. In terms of religion, it was dramatically different from the average with much lower proportions of protestant United Church adherents and Catholics and a very high proportion of Mormons.

The provincial census tracts that were the most similar demographically to the benchmark area were all relatively close to the Twin Butte area of southwest Alberta. This circumstance is simply a reflection of Alberta's physical, economic, and cultural geography. That these areas emerged from the analysis in spite of the exclusion of Mormon religion from the calculation of social distance scores attests to the relatively unique sociodemographic characteristics of the Mormon population and the robustness of the technique. The four census tracts that were the most similar demographically to the benchmark area were deemed the best candidates for the most comparable unexposed reference area. Demographically, there was little on which to base a choice between them, and none of them were exposed to sour gas contaminants. Consequently, Stirling/Raymond, which had the lowest social distance score (ie, 1.4) and was within daily commuting range of the index area, was selected as the preferred most comparable unexposed reference area. The sociodemographic similarity of Stirling/Raymond and the benchmark area is illustrated for all 12 variables in table 2. Only the percentage of the labor force engaged in farming occupations differed between the two areas, a result of the presence of more rural service centers in the Stirling/Raymond area.

In contrast to the selection of the most comparable unexposed reference area identification of a suitable area for the most comparable exposed reference area was much more problematic. Indeed, the areas with social distance scores of less than 15 and the areas exposed to emission from sour gas plants were spatially independent of each other. The distance scores for the areas most demographically similar to the benchmark area and exposed to sour gas were minimally in the 15-to-20 range, as compared to scores of less than four for areas that were demographically similar and not exposed. In our experience with the Alberta results social distance scores higher than 20 render the communities impossible to compare even with statistical adjustments. Social distance scores of less than 15 are workable, and those under 10 are easily handled with adjustments. Scores under four permit comparisons without statistical adjustments, especially as one approaches zero.

Table 2. Comparison of all areas in the study, the benchmark area, and the Stirling/Raymond area — 1981 census data.

Variables ^a	All areas in the study		Benchmark area	Stirling/Raymond area
	Mean	Standard deviation		
Pop 0 to 4	8.6	1.9	11.3	12.8
Pop 5 to 14	16.8	3.0	21.9	22.3
Pop ≥ 65	10.2	4.5	7.8	9.0
6+ families	6.1	3.4	17.2	16.1
Mean income ^b	25 520	4 815	24 555	22 935
Low educ	18.6	6.7	18.3	21.0
Some univ	11.9	4.6	15.6	15.2
Farmers	20.5	16.8	50.0	29.7
Primary	2.7	2.4	.75	0.0
United Church	27.0	8.4	10.7	7.5
Catholic	21.3	7.7	11.6	11.0
Occup ≤ 2 years	37.1	10.3	26.6	28.5
Mormons, 1971	3.4	7.4	40.7	41.6

^a See table 1 for a complete explanation of the variables.

^b In Canadian dollars (CAD).

The five areas exposed to sour gas emissions with the lowest social distance scores all had a much higher proportion of non-Mormon religion, a dramatically smaller family size, and fewer children than the benchmark area. They were all relatively close to Calgary although they differed somewhat in demographic composition according to their proximity to the Calgary metropolitan area. Given these circumstances it was determined that areas that were demographically comparable to the index area and exposed to sour gas emissions do not exist. Faced with this evidence we had to change the main research plan. The main comparison for the health analysis continued to be between the index area and the most comparable unexposed reference area. However, in addition, the social distance algorithm was used to identify an area that was exposed to sour gas emissions and was the most representative of *all* nonmetropolitan southern Albertans (the mean values in table 2). Of the various alternatives, an enlarged area surrounding several gas plants north of Calgary was judged to be the most representative exposed reference area. Thus, for many of the comparisons in the health analysis, we had a sample of "exposed average nonmetropolitan southern Albertans" to compare with a subsample selected from the population of the most comparable unexposed reference area (Stirling/Raymond) by excluding Mormons of that area in many of the comparisons done in the analysis.

Evaluation of the appropriate selection of reference areas can be determined from the actual distribution of relevant demographic variables in the samples drawn for health surveys and medical examination. The similarity in most variables between the index area and the most comparable unexposed reference area is striking (table 3). The differences in proportions and means are small. For some of the continuous variables, small differences attained statistical significance because of the very large sample size we employed. However, these are not of important consequence. With respect to the mean age of the population, we note that the index

Table 3. Comparison of the index area and the Stirling/Raymond area with respect to selected sociodemographic variables for adults 15 years and older — 1985 questionnaire data.

Variable	Index area	Stirling/Raymond area
Sample size (N) ^a	1 396	494
Gender female (%)	48.8	46.6
Age		
15—64 years (%)	85.4	91.1
≥ 65 years (%)	14.6	8.9
Mean age (years)	42.1	39.2
Family income ^b		
< CAD 9 000 (%)	5.2	5.9
CAD 9 000—20 000 (%)	19.4	14.8
> CAD 20 000 (%)	58.9	60.3
Education ^b		
≤ 6 years	2.2	2.2
7 years to some postsecondary	74.5	69.6
Postsecondary and higher	22.8	24.9
Occupation farmer (%)	27.4	24.7
Median length of residence (years)	12.0	9.0
Religion ^c		
United Church (%)	14.3	4.1
Catholic (%)	14.3	16.9
Mormon (%)	50.1	44.3
Family size ^c		
1 person (%)	5.1	3.6
2—5 people (%)	64.3	59.6
≥ 6 people (%)	30.5	30.5

^a A complete survey was undertaken of all residents in the index area, whereas for the Stirling/Raymond area a complete census of all residents was done from which a probability sample, stratified by age and religion, was drawn. In part, this latter procedure was used to obtain a sufficient sample of average nonmetropolitan southern Albertans from Stirling/Raymond who could be compared with a similar sample from the most representative exposed reference area. The town of Raymond was excluded from the survey so that Stirling/Raymond would more closely approximate the rural character of the index area. Specific details are available in the report by Spitzer (20).

^b Percentage for those reporting.

^c Data based on the entire sample of adults and children (index area, N = 2161; Stirling/Raymond area, N = 840).

area was somewhat older and would have tended to produce health status findings that were more adverse for the index area than for the Stirling/Raymond area. In the subsequent health evaluation the data were analyzed with and without statistical adjustment for possible confounders, including sociodemographic variables. The difference for the most important target outcome variables was negligible if one compares the crude rates with those statistically adjusted. Thus, the main results were presented unadjusted, exactly as they were obtained in the field.

Discussion

The objective of this study was to select two reference areas for comparison with the Twin Butte index area in southwest Alberta, ie, the most comparable unexposed reference area and the most comparable exposed reference area. Measurement of demographic similarity was based on 1981 Canadian census data and the cal-

culation of Euclidian distance scores between a specifically defined benchmark area and the other 119 provincial census tracts included in the analysis.

The results indicated a clear choice for the most comparable unexposed reference area. Fortunately, the people of the Stirling/Raymond area accepted the role of principal control community in the main environmental epidemiology project. For a variety of reasons, selection of the most comparable exposed reference area was not feasible. The basic problem was that no other area in Alberta south of Edmonton with sour gas emission was very similar demographically to the index area. This circumstance was due primarily to the unique cultural characteristics of the index area and adjacent parts of southern Alberta — a relatively high proportion of Mormons and associated characteristics of a large family size and high educational achievement. Thus, the research design was altered to select an area exposed to sour gas emissions and representative of all nonmetropolitan southern Albertans.

The Euclidian distance measure has also been used in Ontario for a number of health studies in which the research designs required the identification of reference areas (21 and a report submitted by Murdie to the Ontario Ministry of Health in 1985, a report submitted by Murdie to the Toronto Department of Public Health in 1985, and a report submitted by Spitzer to the Toronto Medical Officer in 1984). Of these, the Toronto Junction Triangle Study (21 and report by Spitzer to the Toronto Medical Officer in 1984) is the most relevant because it is the only study, aside from the Alberta project, for which it is possible at present to evaluate the validity of the procedure. The results from the questionnaires that were administered to a sample of residents in both the Junction Triangle and the comparison census tract revealed a close correspondence between the areas in terms of demographic composition (table 4). The only major difference between the two areas was in ethnic makeup, Portuguese rather than Italian being the major non-Anglo-Saxon group in the Junction Triangle tract. However, this difference had been expected beforehand, and it was subsequently confirmed that it did not prejudice the relevance of the research results.

Although the procedure outlined in this study has been used with considerable success in epidemiologic research, there are numerous methodological issues connected with the application of cluster analysis generally and with the application of social distance measures more specifically. The most important for the present analysis were the measure of similarity, scale of measurement, and the orthogonality issue.

For the present analysis squared Euclidian distance (D^2) was used as a measure of social distance. However, several measures have been suggested for determining similarity (16, 19). Some, such as the chi-square or phi-squared technique, are more appropriate for use with binary data than with continuous data, while others, such as the Pearson correlation coeffi-

cient, measure pattern similarity rather than distance per se. For the application described in this paper Euclidian distance is the preferred measure of similarity. With the Euclidian distance measure itself, controversy concerns the use of D or D^2 . For the present application the controversy is irrelevant since the ultimate purpose is to rank observations according to their similarity with an index area. However, there is a slight advantage in using D^2 in that differences between observations are exaggerated exponentially and it may be easier to identify breaks between groups of observations that are relatively similar to the index area.

Values obtained from Euclidian distances depend on the scale of measurement used in the analysis. For this analysis the 12 variables listed in table 1 were standardized with the use of Z scores (mean=0, variance=1) in order to give each characteristic the same importance in the analysis. This procedure seems reasonable when, as in this study, variables are expressed in different units of measurement (eg, percentage of the population 0 to 4 years of age and average household income) and display very different means and variances. Johnston (12) has, however, argued that, depending on the values of means and variances, there may be situations where the results differ greatly from expectations based on the nonstandardized data.

The third major issue concerns orthogonality. In this analysis the first 12 variables from table 1 were used in standardized form to compute D^2 . Many researchers (13) argue that, since the Pythagorean theorem applies to right-angled triangles, the variables used in the calculation of distance measures should be orthogonal or uncorrelated with each other. A common method of achieving orthogonality is to use scores from a principal components or factor analysis of the data as input to the distance calculations. Other authors (4, 18) see no objection to using correlated variables. Intuitively, this position is justifiable if the correlations between variables are not excessive (excessive defined as $>|0.8|$). The troubling issue then is double counting. It could be argued, for example, that, if measures of income, occupational status, and educational achievement are all relatively highly related and included in an analysis, the concept of economic status gets counted three times. Conversely, it could be argued that each of these variables is conceptually independent and all three should be included in the analysis as individual variables. In practical terms we have found relatively little difference between social distance scores based on the original variables and on an orthogonal transformation of the variables.

The procedure described in this report is an efficient and accurate means of selecting reference populations for epidemiologic studies in which the objective is to maximize demographic similarity between a geographically delineated index area and a suitable reference area. Given the widespread availability of computer-readable census data and the recent inclusion of cluster

Table 4. Comparison of the Junction Triangle area and the reference area (city of Toronto) with respect to selected sociodemographic variables for adults 15 years and over — 1983–1984 questionnaire data.

Variable	Junction Triangle area	Control area
Sample size	279	261
Gender female (%)	51.6	50.2
Mean age (years)	39.4	41.4
Family income ^a		
<CAD 9 000 (%)	13.2	10.2
≥CAD 20 000 (%)	56.6	54.4
Education ^a		
≤6 years (%)	19.6	23.0
7 years to some post-secondary (%)	66.2	63.0
Postsecondary or higher (%)	14.2	14.0
Median length of residence (years)	7.0	9.0
Ethnicity		
Anglo-Saxon	33.7	41.4
Italian	5.4	16.1
Portuguese	14.0	.4
Other	47.0	42.2

^a Percentage of those reporting.

routines with Euclidian distance measures in the standard statistical software packages, the procedure is also relatively easy to operationalize.

Acknowledgments

This study was funded in part by a grant from the Acid Deposition Research Program of the Department of Community and Occupational Health of Alberta, Canada.

References

1. Alberta Environment. Industrial sulphur emissions for Alberta 1977–81. Alberta Environment Pollution Control Division Air Quality Branch, Edmonton 1982.
2. Berry BJL. Grouping and regionalizing: An approach to the problem using multivariate analysis. In: Garrison WL, Marble DF, ed. Quantitative geography, part 1. Economic and cultural topics. Northwestern University, Evanston, IL 1967, pp 219–251. (Northwestern University studies in geography no 13).
3. Berry BJL, Ray DM. Multivariate socio-economic regionalization: A pilot study in central Canada. In: Rymes T, Ostry S, ed. Regional statistical studies. University of Toronto Press, Toronto 1966.
4. Cronbach LJ, Gleser GC. Assessing similarity between profiles. *Psychol Bull* 50 (1953) 456–473.
5. Davies WKD. A multivariate description of Calgary's community areas. In: Barr B, ed. Calgary. University of Victoria, Victoria, BC 1975, pp 231–269. (Western geographical series, no 11).
6. Davies WKD. The social taxonomy of Edmonton's community areas in 1971. In: Smith P, ed. Edmonton. University of Victoria, Victoria, BC 1978, pp 169–197. (Western geographical series no 15).
7. Delta Projects Limited. Alberta/BC gas plant map. Delta Projects Limited, Calgary 1985.

8. Dixon WJ. BMDP statistical software. University of California Press, Berkeley, CA 1985, pp 456—463.
9. Energy Resources Conservation Board. Inventory of gas processing plants in Alberta. Energy Resources Conservation Board, Calgary 1985.
10. Goldberg MA, Mercer J. The myth of the North American city. University of British Columbia Press, Vancouver 1986, pp 226—245.
11. Green PE, Frank RE, Robinson PJ. Cluster analysis in test market selection. *Manage Sci* 13 (1967) B387—B400.
12. Johnston RJ. Grouping and regionalizing: Some methodological and technical observations. *Econ Geogr* 46 (1970) 293—305.
13. Johnston RJ. Multivariate statistical analysis in geography. Longman, London 1978.
14. Norusis, MJ. SPSS^x advanced statistics guide. McGraw-Hill, New York, NY 1985, pp 166—191, 462—467.
15. Ray DM, Murdie RA. Canadian and American urban dimensions. In: Berry BJL ed. *City classification handbook*. John Wiley & Sons, New York, NY 1972, pp 181—210.
16. Romesburg HC. *Cluster analysis for researchers*. Lifetime Learning Publications, Belmont, CA 1984.
17. SAS Institute Inc. SAS user's guide: Statistics, version. Fifth edition. SAS Institute Inc, Cary, NC 1985, pp 255—316.
18. Sokal RR. Distance as a measure of taxonomic similarity. *Syst Zool* 10 (1961) 70—79.
19. Sokal RR, Sneath PHA. *Principles of numerical taxonomy*. Freeman, San Francisco, CA 1963.
20. Spitzer WO. Southwest Alberta medical diagnostic review report. McGill University, Montreal 1986.
21. Spitzer WO, Suissa S, Eastridge L, Shenker S, Germanson T, Murdie RA, MacPherson AS. The Toronto Junction Triangle study: A response to a community health emergency. *Can J Public Health* 77 (1986) 257—262.
22. Statistics Canada. 1981 census dictionary. Ottawa 1982, pp 93—106. (Catalogue 99—901).
23. Veldman DJ. Fortran programming for the behavioral sciences. Holt, Rinehart and Winston, New York, NY 1967, pp 308—317.
24. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58 (1963) 236—244.

Received for publication: 16 March 1988