SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853

TECHNICAL REPORT NO. 840

March 1989

MONITORING FOR CLUSTERS OF DISEASE; APPLICATION TO LEUKEMIA INCIDENCE IN UPSTATE NEW YORK

by

Bruce W. Turnbull, Ph.D. Cornell University
Eric J. Iwano, M.S., Mount Union College
William S. Burnett, M.D., New York State Dept. of Health
Holly L. Howe, Ph.D., Illinois Dept. of Public Health
Larry C. Clark, Ph.D., University of Arizona

Address reprint requests to: Bruce W. Turnbull, 334 Upson Hall, Cornell University, Ithaca, NY 14853

Support for the first two authors was provided in part from Grant #GM 28364 from U.S. National Institutes of Health.

Running title: MONITORING FOR DISEASE CLUSTERS

MONITORING FOR CLUSTERS OF DISEASE; APPLICATION TO LEUKEMIA INCIDENCE IN UPSTATE NEW YORK

Bruce W. Turnbull, Ph.D., Cornell University
Eric J. Iwano, M.S., Mount Union College
William S. Burnett, M.D., New York State Department of Health
Holly L. Howe, Ph.D., Illinois Department of Public Health
Larry C. Clark, Ph.D., University of Arizona

ABSTRACT

We propose a procedure for the detection of significant clusters of chronic diseases, with particular reference to cancer. The procedure allows for variations in population density and avoids the problem of "post-hoc" formation of hypotheses or self-defined populations. This accounts for several of the principal problems of cluster evaluations.

The procedure defines a set of overlapping "windows" or areas of constant population size (n=2500, for example) centered on an irregular grid formed by the centroids of Census block groups. The adjusted incidence rates are calculated for each window. The (very large) number of resulting rates can be regarded as identically distributed but not independent random variates. The distribution of the extremes of this spatial process under a null hypothesis of randomness can be obtained by permutation or Monte Carlo methods. The whole procedure is then repeated for a succession of increasing window area population values. The significance level of any observed extreme area disease rate or observed pattern of disease case incidence can then be calculated from the computed null distributions. The techniques are practical but "computer-intensive". The procedure, termed the "Cluster Evaluation Permutation Procedure" (CEPP), is applied to leukemia incidence data for an Upstate New York region obtained from the Cancer Registry and Census files. We also make comparisons with two other recent clustering methods proposed by Whittemore et al. (1) and by Openshaw et al. (2). Routine examination of disease occurrence with CEPP would allow state health officials to prioritize case investigations and to respond in a timely and efficient manner to inquiries of reported clusters.

1. Introduction

In this paper we discuss statistical methodology for the definition and detection of significant clusters of chronic diseases with particular reference to cancer. This methodology could be used as part of a proactive disease surveillance system by public health departments. The study of the possible existence of clusters of cancer has a long history (3,4) yet the subject continues to stimulate much publicity and controversy. The cases of Woburn, Massachussets and Dounreay, Scotland are but two of many examples. The need for a systematic and objective way to detect, prioritize and monitor the occurrence of statistically significant clusters of disease is apparent. State Departments of Public Health expend sizeable resources investigating reported clusters, only a few of which eventually turn out to be of justifiable concern. In a survey of cancer cluster procedures of State Health Departments, Warner and Aldrich (5) found that "cluster investigations had generally been unproductive in terms of etiologic discoveries, yet may have important benefits in terms of public education, allaying public anxiety about environmental concerns and engendering good will toward government agencies". (See also the remarks by Rothman (6) and by Schulte et al. (7).) Statistical methods for assessing overall patterns of disease incidence should be a part of a proactive surveillance program. With the results of such a program in hand, it will be easier to respond to lay reports of perceived clusters in a timely manner.

For initial investigation, the eight contiguous counties in upstate New York, Broome, Cayuga, Chenango, Cortland, Madison, Onondaga, Tioga, Tompkins, bordering on and including Cortland County, were selected as "Study Region A". (See Figure 1.) These are eight of the nine counties that make up Region 7 as defined by the New York State Department of Environmental Conservation. (Geocoded data for the ninth county, Oswego, were unavailable at the time of this analysis.) This area of Central New York of over one million people consists of rural communities primarily but it does contain two large metropolitan areas, Syracuse and Binghamton, and several smaller cities, including Cortland and Ithaca.

The initial disease to be studied is leukemia with the hopes that methods generated for its surveillance will also be applicable to other forms of cancer. Leukemia was selected as the "model" cancer because of its remarkably uniform distribution although apparent clusters have been known to occur. Some of these may

have been the result of point source environmental exposures such as might have occurred in Woburn, Massachusetts (8,9).

The study region was divided up into 713 "cells". For Broome County the cells were defined as the 55 U.S. Census tracts. For the other seven counties the smaller Census block group units were used. (Typically a census tract contains 1000-4000 people and comprises of between two and five block groups.) Demographic data, including population sizes, were obtained for each cell in the study region using the 1980 U.S. Census. Centroids of the block groups are displayed in Figure 1. The denser parts on the map reflect the more urban areas. In cooperation with the New York State Cancer Registry, all 592 cases of leukemia in the study region reported during the five year period 1978-1982 were geocoded and placed in the cells. (For Broome County the geographic location of cases was available only at the larger census tract level; more precise block group placement was possible for the other seven counties.) The locations of these 592 cases are shown on Figure 2. Upon comparing Figures 1 and 2, it becomes obvious that there is a need for taking into account varying population density when assessing visual appearance of clusters on a map such as Figure 2. Density equalized map projections or cartograms (10), i.e. maps with land areas distorted to be proportional to population size, can assist the presentation of the case residence data. For such a map of New York State, see Figure 2 of Levison and Haddon (11). However such maps do not address directly the statistical issue of significance of case clustering.

It should be noted that some of the cases could be geocoded only to one of two or three possible block groups due to ambiguities in the address records. Approximately 90% of the cases were geocoded precisely to the block group level and 95% to the census tract level. The incomplete data were handled in two ways. In the first, such cases were assigned fractionally to the possible block groups proportional to their corresponding population sizes. In the second way, such cases were also fractionally assigned, but now proportional to the number of cases observed in those areas. The first way tends to make the distribution of cases seem less clustered, the second the opposite. All analyses were performed using both methods of handling the incomplete geocoding. However little difference was found between the two methods, and thus only results using the former method are reported. Of course there might be differences in a situation in which the geocoding was less complete.

In this paper we will describe a "cluster evaluation permutation procedure" (CEPP) for analyzing spatial patterns with varying population density. The method is compared to two other recently proposed procedures, the U-statistic test of Whittemore et al. (1) and the "Geographical Analysis Machine" of Openshaw et al. (2). It should be noted that all three methods can accommodate stratified analyses; that is, where area rates are adjusted by age, sex and other demographic variables that are available in the Census and Cancer Registry data. However, for the sake of brevity and clarity of exposition, we describe only the unstratified analyses which are appropriate for populations that are homogeneous demographically. The Upstate New York leukemia data are used only to illustrate and compare the methodologies; the results should not be interpreted as a full analysis of the data. Such an analysis would include consideration of histologic type of the disease, adjustments for stratifying and concomitant variables such as age and sex, and a more critical examination of the geocoded case data, which might include tracing residence histories of the cases.

2. The U-Statistic Test of Whittemore et al.

Suppose we wish to test a "randomness" null hypothesis H that each member of the study population is equally likely to be a "case". Whittemore et al. (1) describe a test for the detection of clusters that is based on the mean distance between all pairs of cases. The variance and expected value of the test statistic are computed using underlying population data. Under the null hypothesis, the numbers of cancer cases in each cell of the study region are treated as independent Poisson random variables with means proportional to the populations of the cells. The authors prove that their test statistic is asymptotically normally distributed under the null hypothesis of random placement of the cases. The authors also present a stratified version of their statistic. For our Upstate New York leukemia data, the observed mean distance between all pairs of cases was 60.24 km., as compared with an expected value of 59.01 km. and standard deviation of 0.96 km. Thus Z = 1.29 which is not significant.

There are however two drawbacks of the method of Whittemore et al. (1). First, the method does not indicate the position of significant clusters, if one or more exist. Second and perhaps more serious, although its null distribution does depend

on the population density pattern, their test statistic depends on the position of the cases only through their pairwise distances. Thus, for a given population distribution, the test cannot differentiate between a situation with an apparent cluster in an urban area (probably due to high population density) and one in a rural area (high rate) when the pairwise distances are the same. The statistic will tend to be negative if a cluster exists in an urban area, (many short arcs) but positive if there is a cluster in a rural area (many long arcs). In particular the statistic has poor power for detecting a cluster in a medium town. These intuitive remarks are supported by simulation results (12).

3. The Geographical Analysis Machine of Openshaw et al.

Openshaw et al. (2) take a graphical approach, using what they term a "Geographical Analysis Machine" (GAM). The method examines overlapping circular areas and notes those with high rates. Their algorithm proceeds as follows:

1. Select a radius r, e.g. 1, 2, or 4 km.
2. Lay down a closely spaced square lattice over the study region with grid points even spaced at intervals r/5, say, apart. Label the grid points $i = 1,2,.....,K(r)$.
3. Consider each of the K(r) grid points in turn. For the i'th grid point compute C_{ir}, the number of cases in a circle with radius r and centered at that grid point. Draw in the circumference of that circle if the observed value of C_{ir} is two or more and exceeds the 99.8'th percentile of the distribution of the number of cases to be found in this circle under the null hypothesis.
4. Now return to Step 1 and repeat the procedure for the next higher value of r.

The result is a map covered with a number of circles. It should be noted that because of the high degree of overlap, there is considerable correlation between the C_{ir} values of neighboring circles. Hence the circumferences that are drawn will appear bunched densely together in clusters even under the null hypothesis. However the drawn circles will become more bunched if clustering is present.

The procedure is computer intensive because there may be several hundred thousand grid points and four or five values of r. For each combination, both the

value of C_{ir} and the 99.8 percentile cutoff point must be computed. In fact the method is even more highly computer intensive because the authors propose using Monte Carlo simulation to obtain these 99.8 percentile values. This involves generating 499 replications of the spatial point process under the null hypothesis that each of the N members of the population of the region is equally likely to be one of the C cases, where C is the total number of cases in the study region. For each replication the steps 1-4 are repeated. It is no surprise that the authors needed a Cray XMP "supercomputer" to analyze their data set which comprised of C = 853 diagnosed cases of acute lymphoblastic leukemia in a population of N = 1,544,963 children in Northern England. We used a less computationally intensive method for obtaining the 99.8% cutoff value for our Upstate New York data set. Instead we used the 99.8'th percentile of a Poisson distribution with mean $\mu = C^{*}P_{ir}/N$ where P_{ir} is the population contained in the circle of radius r centered at the i'th grid point. This drastically reduces the computing load by eliminating the need to search 499 simulated case vectors for each grid point, but even so the load is still considerable: with the three values of r, there were 83,587 circles to be examined ! The results for radii of 1,2, and 4 km. are displayed in Figure 3.

The GAM procedure provides an excellent descriptive method for finding areas with high rates that are free of geopolitical boundaries. However, before coming to any conclusion on existence of clusters based on Figure 3, it should be noted that the method does not lead to a quantitative assessment of significance of an observed pattern. The .002 "significance" level appears to have been chosen by convenience so that a reasonable number of circles appear on the maps. As mentioned above, because of high correlation between overlapping circles, there will be some apparent clusters even under a pure randomness assumption. Thus, although the Figure perhaps indicates the existence of some clusters, it is not clear whether the method is just picking out clusters which must occur at some locations just by chance (11,13,14). Simulations can reveal what patterns of "significant" circles typically occur under the null hypothesis, but since these depend on the disease rate, the area and the population density structure, the entire exercise would have to be repeated for each new study.

4. Cluster Evaluation Permutation Procedure

The Cluster Evaluation Permutation Procedure is a procedure for defining and detecting the presence of clusters and for assessing their significance, i.e. how likely they are to be spurious due solely to chance.

The study region is divided up into a large fixed number I, say, of cells, typically census tracts or block groups. The distance between two cells is defined to be that between the geographical centroids of the two cells. We define N_i to be the population of the i'th cell, and let $N = \sum N_i$ be the total population in the region. Also let C_i be the number of cases in the i'th cell, usually 0 or 1. Of course $\sum C_i = C$, the total number of cases. For each cell i, we form a "window" or two-dimensional "ball" of neighboring cells so that its population is R, where R is a fixed number of persons, e.g. 2500. More precisely the ball is formed as follows. We consider each cell in turn. For cell i, assuming that its population N_i is less than R, we examine the cell whose centroid is closest to that of cell i, say this is cell j. If $N_i+N_j = R$, then cell j is included along with cell i, and the formation of the i'th ball is completed. If $N_i+N_j > R$, then we take only a "fraction", $(R-N_i)/N_j$ of cell j, completing the formation of ball i. If $N_i+N_j < R$, this cell is fully absorbed into ball i and we continue the process by examining that cell whose centroid is the next closest to that of cell i, etc. Hence ball i contains cell i and a collection of its nearest neighboring cells, the furthest one being only "fractionally" represented. If $N_i>R$ then the ball consists only of a fraction of cell i; however in our example R was chosen sufficiently large that this did not happen. We now have a collection of I overlapping balls each with constant population R. For ball i, we now compute C_{iR}, the number of cases occurring in that ball. This is the sum of the numbers of cases in cells totally included in that ball plus the corresponding fraction of those cases in the cell only partially included. (The values C_{iR} should not be confused with those of the procedure of Openshaw et al.(2). Their values were based on equally spaced grid points and circular areas of constant geographical radius.) Because the balls are based on areas of equal population, the $\{C_{iR}\}$ can be viewed as directly proportional to the disease rates. Under the null hypothesis of randomness, the values $\{C_{iR};\ i=1,2,\ldots,I\}$ can be considered as identically distributed (but not independent) random variables, and hence can be used to test this hypothesis.

Various test statistics can be constructed from the $\{C_{iR};\ i=1,2,\ldots,I\}$ that will be sensitive to departures from the null hypothesis. A natural choice is the extreme value or maximum statistic $M_R = \max(C_{1R}, C_{2R}, \ldots\ldots, C_{IR})$, as this might trigger an alarm to a state health department. A statistical test can be based on the null distribution of M_R. The null hypothesis is rejected if M_R is greater than some cutoff value K, where K is determined by the null distribution of M_R and the significance level α. Alternatively the P-value is given by the probability, computed under the null hypothesis, that M_R exceeds its observed value. We also note the identity of the cell that corresponds to this maximum M_R value.

The null distribution of M_R can be obtained by using the randomization test ideas described by Fisher (15). Here the distribution is the collection of M_R values obtained by considering all ways of assigning the C cases to the N persons in the population, these assignments being equally likely. Usually the number of such permutations, N!/(C!(N-C)!), will be too large to allow the distribution to be computed exactly and it is necessary to take a Monte Carlo sample of the permutations thereby obtaining an estimate the distribution. Monte Carlo tests were proposed by Barnard (16) and their general properties have been investigated by Hope (17) and Marriott (18). Applications to analysis of spatial patterns have been described by several authors, including Ripley (19), Besag and Diggle (20), and Raubertas (21). This idea was also used by Openshaw et al. (2) as described previously. However unlike Openshaw et al. who were looking at a very large number of statistics, the occurrences in each of many circles, we obtain only the maximum statistic, M_R. Thus without the multiplicity effect it is not necessary to estimate such small nominal significance levels (e.g. 0.002) with the resulting need for a large number of Monte Carlo replications. For example with 99 simulations an exact test at level 1% rejects the randomness hypothesis if and only if the observed value of M_R exceeds the largest of the 99 simulated values. Correspondingly the P-value would be the proportion of simulated values that exceed the observed value of M_R.

The results for our Upstate New York leukemia study are shown in Figure 4. Because there is no natural choice for R, four representative values were chosen, namely 2500, 5000, 10000, 20000 persons. The choices for R will depend on the average size of the cells and on the disease and exposure pattern under consideration. For example, if radon, air or water pollution point sources for the

disease are suspected, R could be chosen to be the typical number of people that would be affected. The Monte Carlo simulation involved drawing 999 replications from a multinomial distribution with C independent trials and N mutually exclusive outcomes whose probabilities are each equal to 1/N. For our study, the population is N=1,057,673 persons and the number of cases is C = 592. Figure 4 shows the observed values of M_R for the four values of R along with the upper one- and five-percentiles , and the lower five-percentiles of the simulated distribution. As can be seen none of the observed values of M_R are significant at the 1% level; for R = 20,000, it is just significant at the 5% level. However, since we are using 4 values of R, a Bonferroni adjustment (22, page 8) might be used to account for the multiplicity of hypothesis tests, in which case the significance is even further reduced (by a factor of 4). As might be expected because of correlation between overlapping balls, the M_R value for three of the four values of R come from the same block group (M1), and the ball, centered on block group (H), that corresponds to M_R for R=20000 includes that same block group (M1) although it is not the "center" of it. Both block groups are in the cluster in the west part of Cortland County in Figure 3 as "identified" by the method of Section 3.

A similar exercise can be performed for looking at the second and third highest, etc. of the area rates C_{iR} $(1 \leq i \leq I)$. These would also be natural to look at, as they might signal alarms to a state health department. The same simulations can be used to get the null distributions for these statistics.

An important issue concerns the power of the cluster evaluation permutation procedure in the presence of one or more clusters. Simulation studies (12) suggest that the procedure described in this section does have good power to detect the presence of one or more clusters. These results will be described in detail in a forthcoming paper.

In a region with many areas, there will be by chance variation a wide variation in disease rates. One area must of necessity have the highest rate. The question is whether this highest rate is unusually large, compared to what the highest rate could be expected to be under purely random variation. This question is the one directly addressed by our method.

Rather than being used as an absolute measure, we might recommend that the P-value be used in a relative way to prioritize areas for cluster investigations, in fact the P-values can be compared even across different diseases. Alternatively, orderings could be based on the ratio or difference of the observed maximum rate M_R/R relative to its expected or median value as computed under the null hypothesis.

5. Conclusion

All three methods described here avoid the problems of testing significance of a cluster in a self-defined population. However the CEPP method described in Section 4 enables the quantitative assessment of the statistical significance of apparent clusters. Furthermore, by considering various population radii, it is possible to address the problem of defining the borders of a cluster. The magnitude and type of cluster can be important for directing future epidemiologic research and focussing on the different exposures which may be related to large clusters or to patterns that suggest more intense point exposures. We intended that this method be developed as a procedure for surveillance of chronic disease occurrence. Routine examination of disease occurrence with CEPP would allow state health officials to prioritize case investigations and to respond in a timely and efficient manner to inquiries of reported clusters.

ACKNOWLEDGMENTS

The authors would like to thank Dr Philip Nasca, Director of the Bureau of Cancer Epidemiology, New York State Department of Health, for his encouragement of this project. They are grateful for helpful advice and comments from Professor Sir David Cox, and also from colleagues at Cornell, Walter Federer, Charles McCulloch, Steven Schwager, Fran Thompson and Lance Waller. The first two authors' research was supported by National Institutes of Health Grant No. GM 28364.

REFERENCES

1. Whittemore, A., Friend, Brown, B.W. and Holly, E.A. (1987) A test to detect clusters of disease. Biometrika, 74, 631-637.

2. Openshaw,S., Charlton,M., Craft,A.W. and Birch, J.M. (1988) Investigation of leukaemia clusters by use of a geographical analysis machine. Lancet (6 Feb. 1988), 272-3.

3. Pearson, K. (1913). Multiple cases of disease in the same house. Biometrika 9, 28-33.

4. Kellet, C.E. (1937). Acute myeloid leukaemia in one of identical twins. Archives of Diseases in Childhood 12, 239-252.

5. Warner, S.C. and Aldrich, T.E. (1988). The status of cancer cluster investigations undertaken by state health departments. American Journal of Public Health, 78, 306-307.

6. Rothman, K.J. (1987). Clustering of disease (Editorial). American Journal of Public Health 77, 13-15.

7. Schulte, P.A., Ehrenberg, R.L. and Singal, M. (1987). Investigation of occupational cancer clusters: Theory and Practice. American Journal of Public Health 77, 52-56.

8. Cutler, J.J., Parker, G.S., Rosen, S., Prenney, B., Healey, R. and Caldwell, G.G. (1986). Childhood leukemia in Woburn, Massachussets. Public Health Reports 101, 201-205.

9. Lagakos, S.W., Wessen, B.W. and Zelen, M. (1986). An analysis of contaminated well water and health effects in Woburn, Massachussets. Journal of the American Statistical Association 81, 583-614.

10. Schulman,J., Selvin,S. and Merrill,D.W. (1988). Density equalized map projections: a method for analysing clustering around a fixed point. Statistics

in Medicine 7, 491-506.

11. Levison, M.E. and Haddon, W. (1965). The area adjusted map: An epidemiologic device. Public Health Reports 80, 55-59.

12. Iwano, E. (1989). A comparison of spatial cluster detection procedures. M.S. Thesis. Cornell University.

13. Glass, A.G., Hill, J.A. and Miller, R.W. (1968). Significance of leukemia clusters. Journal of Pediatrics 73, 101-107.

14. Enterline, P.E. (1985). Evaluating Cancer Clusters. American Industrial Hygiene Association Journal 46, B10-13.

15. Fisher, R.A. (1935). The Design of Experiments. Edinburgh: Oliver & Boyd.

16. Barnard, G.A. (1963). Discussion of Professor Bartlett's paper, Journal of the Royal Statistical Society B, 25, 294.

17. Hope, A.C.A. (1968) A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society B, 30, 582-598.

18. Marriott, F.H.C. (1979). Barnard's Monte Carlo tests: How many simulations. Applied Statistics, 28, 75-77.

19. Ripley, B.D. (1977). Modelling spatial patterns. Journal of the Royal Statistical Society B, 39, 172-212.

20. Besag, J, and Diggle, P.J. (1977). Simple Monte Carlo tests for spatial pattern. Applied Statistics, 26, 327-333.

21. Raubertas, R.F.(1988). Spatial and temporal analysis of disease occurrence for detection of clustering. Biometrics, 44, 1121-1129.

22. Miller, R.G. (1981). Simultaneous Statistical Inference. 2nd Ed. New York: Springer-Verlag.

Figure 1

STUDY REGION

Centroids of Census Block Groups

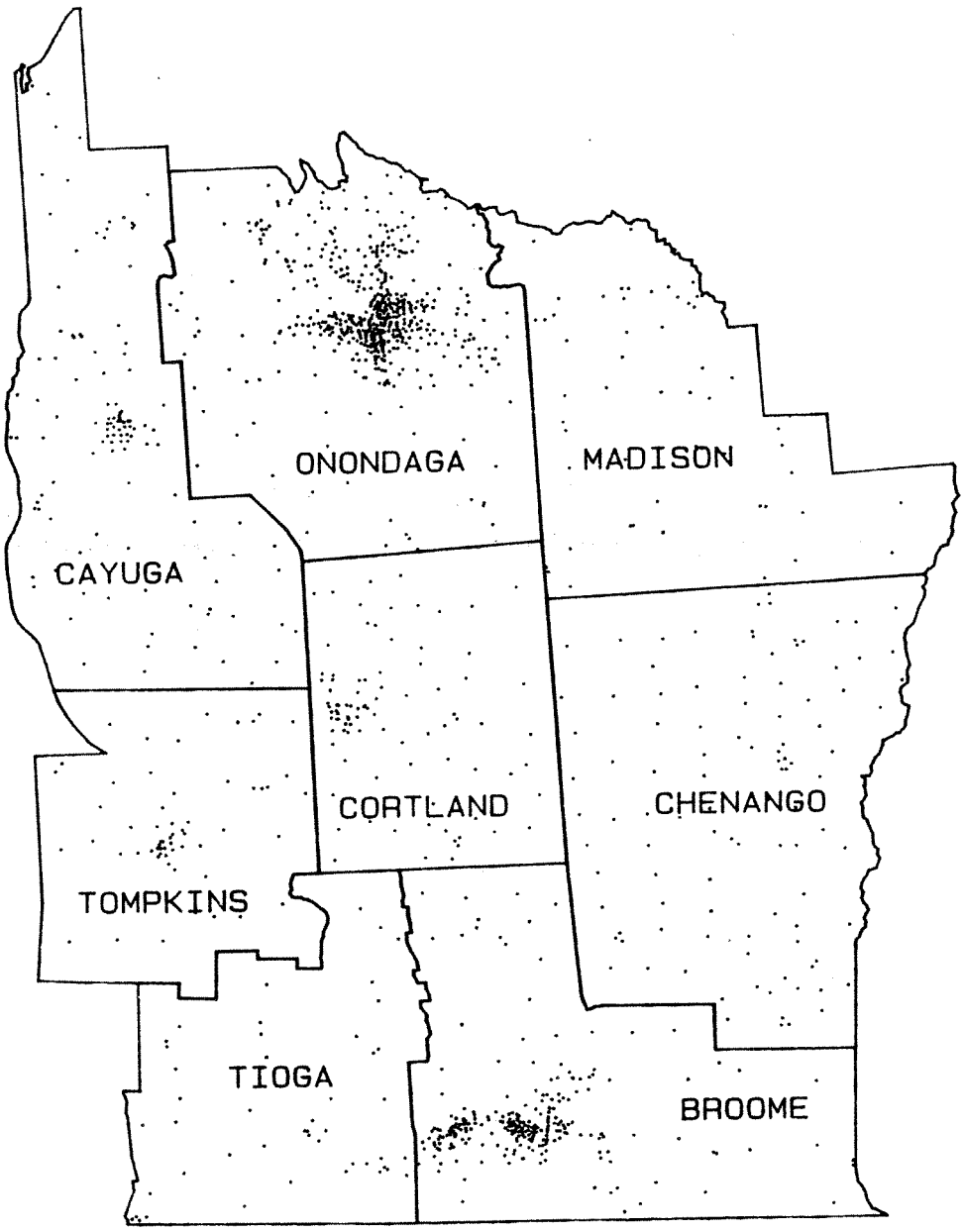

Figure 2

LOCATION OF CASE RESIDENCES

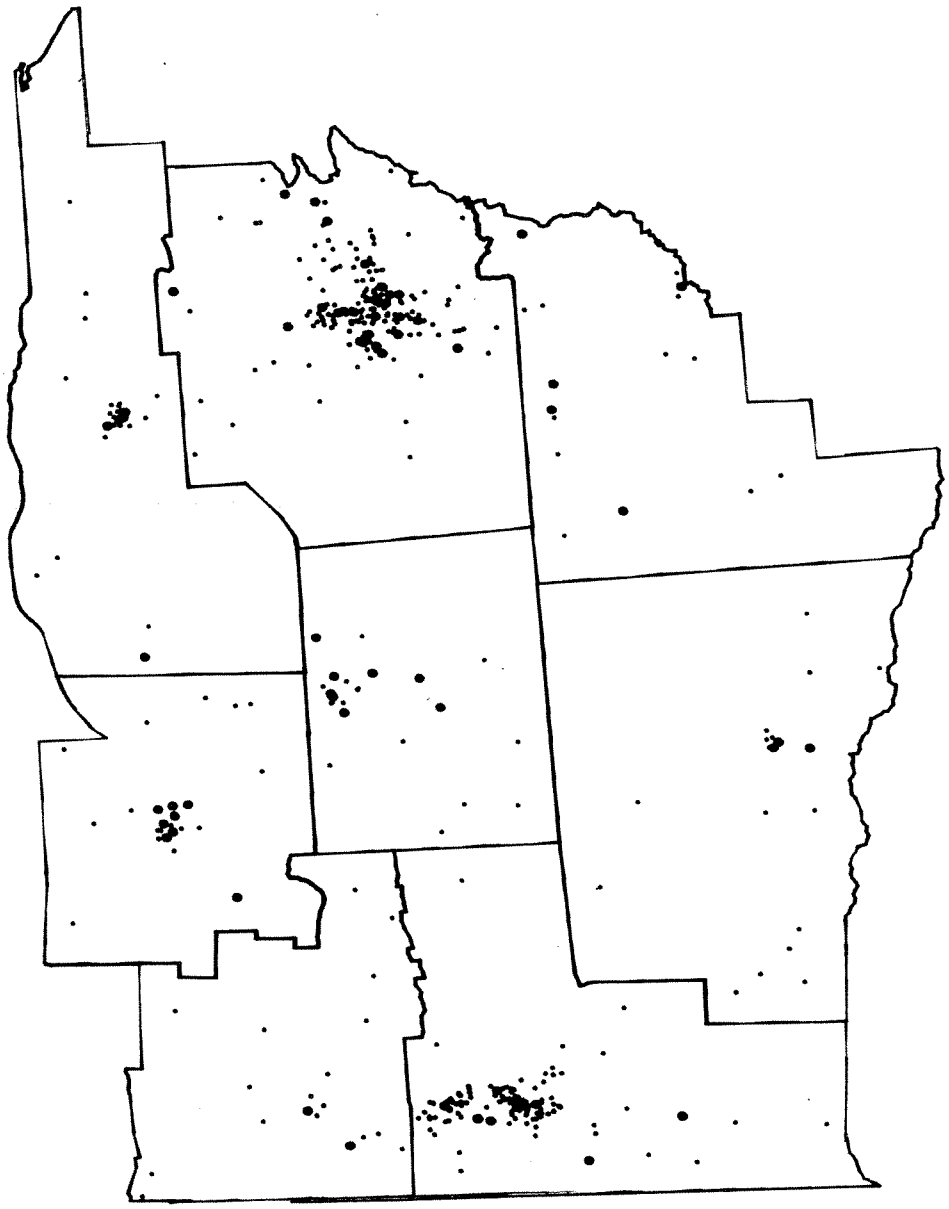

Figure 3

OPENSHAW et al. GAM

1, 2, 4 km Radii

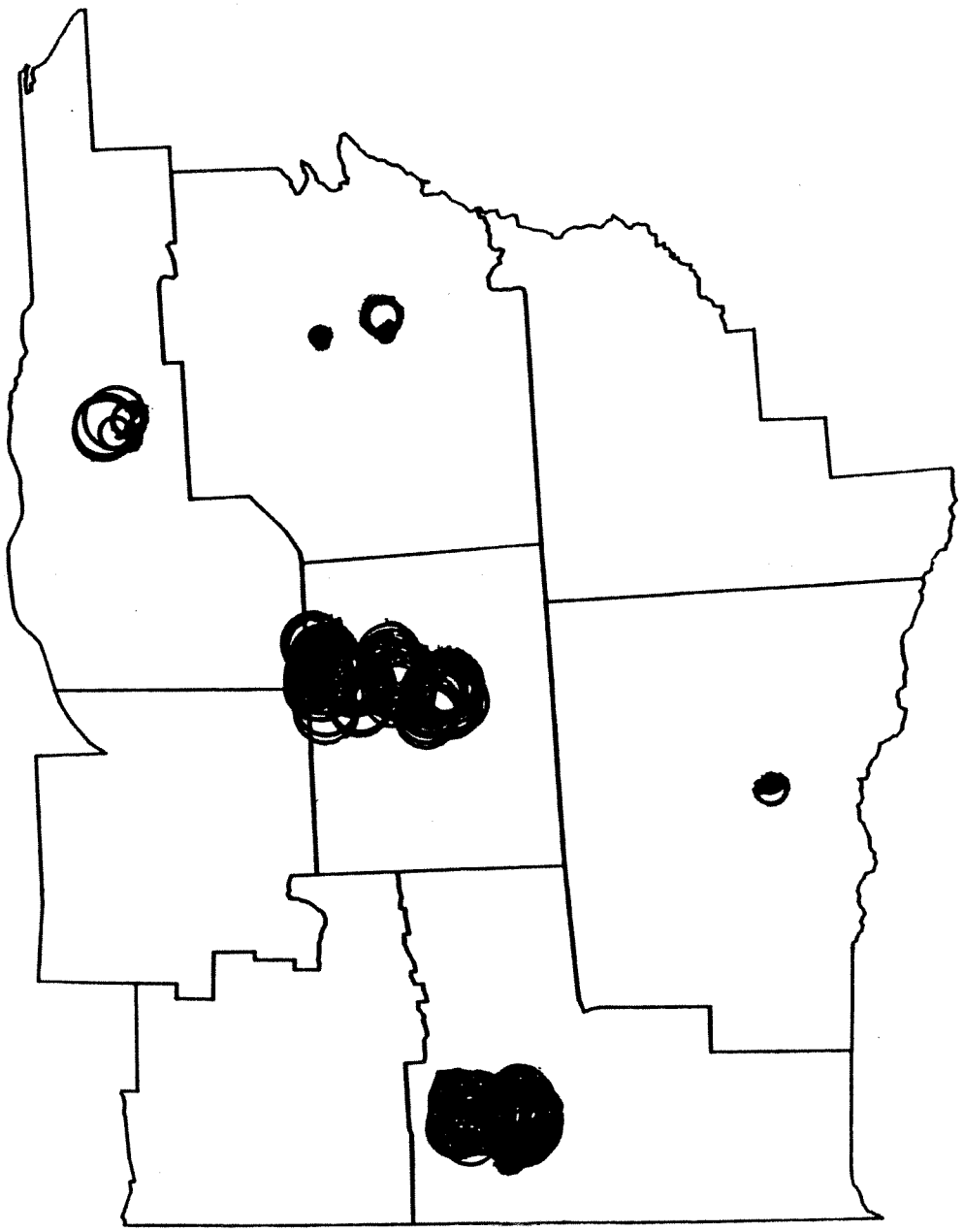

Figure 4

EMPIRICAL SIGNIFICANCE LEVELS

Maximum M_R