THE PROBLEM OF GEOGRAPHIC CONTIGUITY - A MONTE CARLO APPROACH

Stephen F. Gibbens, James A. Tonascia
California State Department of Public Health

This paper describes a method for determining whether patterns of counties, census tracts or other types of geographic units depart from a configuration that might have occurred by chance. The method presented is conceptually simple and can be applied with comparative ease through the use of a computer program available through SHARE[1]. This paper describes the problem, the method of solution and gives some useful probability distributions.

Suppose, for example, a rate of some type was calculated for each of the N counties in a particular state and a map was shaded, showing the N/4 counties having the highest rates. This top quartile of counties might distribute themselves in a number of ways - varying from a single cluster forming a large clump to the opposite situation, where most of the selected counties were isolated, surrounded by counties with low rates. Whatever pattern was formed, most observers would make a judgment on the meaningfulness of the pattern manifested. If the high rate counties tended to be those containing large metropolitan centers, this fact would probably be noted rather quickly. If, however, a judgment was made whether the high rate counties showed a pattern of contiguity - or geographic contagion - the problem might become much more difficult. The point at which a pattern departs from simple randomness is obscure indeed, when it must be based upon visual inspection.

A number of criteria could be formulated defining nonrandomness. A tendency for the chosen counties to form a few clusters, a preponderance of chosen counties falling in a single quadrant of the state, a "least squares" solution based on distance from some focal point - all might be developed into probability sampling distributions for departures from randomness. In this paper, two criteria are developed: (1) the largest cluster formed by the selected counties; and (2) the number of pairs of contiguous counties found among those counties selected. These will be referred to as the "clusters" and "pairs" methods.

Clusters Method

From a shaded map we have a visual picture of geographic contiguity. A nonrandom pattern is characterized by clusterings of shaded counties. However, our data could be more easily quantified if we described the clustering in terms of the number of counties found in the largest cluster formed by the N/4 sample. With this criterion, it is possible to conveniently describe distributions of many samples, and to obtain probabilities of the occurrence of clusters of various sizes. In order to effect this quantification, a listing is required for each state showing each individual county with its contiguous counties.[2]

With the availability of a "county contiguity list" (See Table A), it is possible to select a sample of N/4 counties, from a particular state and note the largest cluster formed by these selected counties.

Briefly described, say we have a state with 100 counties and by use of random numbers select a sample of N/4 counties, notated as A, B, C, . . . Y. We search the county contiguity list for the counties contiguous to A. If counties B and J are contiguous to A, we know we have a cluster of at least three counties. Next, we scan the list of counties contiguous to B and find counties A and F. County A was already counted but F is a new county and so our cluster is known to include at least four counties; A, B, F and J. We proceed similarly to search the lists of counties contiguous to F and J, incrementing the cluster size by one for each sample county found. We continue to county C (the first of the remaining counties sampled that is not already known to be in our first cluster of contiguous counties), repeating the above procedure, and hence on to county Y, the last county included in our sample. When we have finished, we record the number of counties in the largest cluster formed. If we were to repeat this procedure for many samples, accumulating the results, we would have an approximate sampling distribution for the largest cluster size found for that state when N/4 counties are selected by chance. For example, the State of Iowa has 99 counties from which 3,000 independent samples of 25 counties each were selected. In about 5 percent of these 3,000 samples a cluster as large as

[1] The computer program is available through the SHARE users group (identified as GO BC GEOG), with specifications.

[2] It should be noted here that the determination of contiguity or noncontiguity is often quite arbitrary. In this paper, two counties were required to share a common border in order to be considered contiguous. If they were separated by water narrow enough to be bridged, they were considered contiguous. Hence, the Mississippi was not a barrier to contiguity but Lake Michigan was. However, different investigators may want to apply different rules for county connections. For example, when four counties meet at a point, so that their borders form a "cross" (+), a decision must be made as to whether or not the diagonal counties are to be considered contiguous. If "point contiguity" is accepted, i.e., if the diagonal counties are considered contiguous, does this carry over to the situation where the borders describe pairs of obtuse angles, rather than right angles as was the case for the cross? These decisions should be made only after careful consideration of these problems, since alternatives produce significantly different results.

10 counties was found; in 1 percent, a cluster as large as 13 counties was found, and in .1 percent, a cluster as large as 16 counties was found.

Pairs Method[3]

The use of the largest cluster size found is not always an ideal measure of geographic contiguity. If there were several smaller clusters, there might be a departure from randomness even though the largest cluster had only a moderate probability level. Another approach would be to obtain the sampling distribution of the number of pairs of contiguous counties found when selecting many random samples of N/4 counties. A large number of pairs of counties could be found either in the situation where there was a single very large cluster, or, several smaller clusters.

The number of pairs of contiguous counties found in a sample would be ascertained, again by recourse to the county contiguity list, accumulating a tally for every border common to two counties included in the sample. In the above example for the State of Iowa, in 3,000 samples, 19 pairs of counties were found in 3½ percent of the samples, 21 pairs in 1 percent of the samples, and 24 pairs in .1 percent of the samples.

Discussion

The "clusters" method and the "pairs" tend to give similar results when applied to live data. The pairs method is more sensitive to the situation where several clusters of moderate size are formed by the sampled counties, sometimes showing a significant probability level where the clusters method is not significant in terms of the number of contiguous counties in the largest cluster. However, this gain in sensitivity in the pairs method, is in equal measure lost, when contiguity is concentrated in a single cluster. In this situation, the clusters method may show a significant probability level while the pairs method shows a very moderate probability.

Figures 1 and 2 illustrate these different situations. Figure 1 shows the lowest quartile of counties in terms of live birth rate for Iowa in 1962. The largest cluster contains 13 counties. This occurred less than once in every 100 trials when using chance methods. However, among these same counties there were 17 pairs (instances of 2 low quartile counties sharing a common border). This occurred about 12 times in every 100 trials when using chance methods. Figure 2 shows the contrasting situation where the cluster of 8 counties was shown to have occurred over 11 times in every 100 trials whereas the pairs method occurred only twice in every 100 trials.

Since the various states tend to contain different numbers of counties and since the configuration of counties tends to vary as well, clearly, each state has its own unique sampling distribution. Even where two states have the same number of counties and the appearance of the counties is similar, their sampling distributions are often distinctly dissimilar when subjected to a Kolmogorov-Smirnov test. The general appearance of the sizes and shapes of the counties in a state gives little indication of the probability pattern that will emerge.

Table 1 shows the Estimated Cumulative Probability Distribution for Cluster Sizes of Contiguous Counties for Selected States based on 3,000 samples drawn from each state. Each Sample included about 25 percent of the counties in the state. Referring again to the State of Iowa, the table shows that in .001 of the samples, the largest cluster formed by the 25 sample counties was 2 counties; that in about half the samples (.492) the largest cluster found was 5 counties, or less. Similarly, in 95.2 percent, the largest cluster formed was 10 counties or less. The largest cluster of contiguous counties found in the 3,000 samples drawn was 16 counties. From a sample of 25 counties in the State of Iowa, it would be rare indeed to find the largest cluster to be as few as 2 counties, or as many as 16 counties.

Since these probability distributions were developed by the Monte Carlo method, they are qualified as estimated distributions. Barring a bias in the pseudo-random number generator used by the computor (BC RNDY), the true probability limits could be estimated by the usual method.

For a 10-county cluster in the State of Iowa which has an estimated cumulative probability of .952, a .05 level confidence interval would be calculated as:

$$.952 \pm 1.96 \sqrt{\frac{(.048)(.952)}{3{,}000}} = .952 \pm .008$$

There appear to be two main problems in the practical application of this methodology. Let us say an investigator is attempting to ascertain whether a particular type of congenital anomaly found in newborn infants shows a nonrandom geographic distribution. Usually the investigator would not know in what percent of the counties the anomaly should show high rates - perhaps in N/4 of the counties or perhaps N/10. This would depend on the distribution of the factors which he hypothesizes gives rise to high incidence of the anomaly. Since each sample size (N/4 or N/10) would have a different sampling distribution, the task of selecting an appropriate sample size becomes critical. The only solution to this would seem to be to rank the counties according to incidence rates and if some of the high incidence counties were tested to show significant departures from a norm - to cut off below these counties and use that number as the sample size. The use of "fortuitous" distributions would not appear to invalidate the ultimate probabilities since these are based on the geographic pattern manifested, rather than on the statistical significance of the rates in the selected counties.

[3] The authors are indebted to Donald Loveland for suggesting this approach.

The second problem is more subtle. The probability distribution obtained is valid and useful in an a priori sense. However, usually an investigator has the data on a shaded map and is interested in knowing whether the distribution is nonrandom. When viewed in an a posteriori sense, it may appear invalid to use the tabled probabilities because of gross irregularities in the particular counties included in the shaded portion of the map. For example, if a shaded map of California showed a number of the coastal counties in the high rate quartile, the tabled probabilities would be inappropriate for comparison because the probabilities were obtained from all counties. Since the coastal counties have no contiguous counties on their ocean side, they would have less opportunity to be included in large clusters and no opportunity to contribute pairs for a portion of their borders. Therefore, clusters of counties which included coastal counties would occur less often than suggested by the tabled probabilities. At the other extreme, an unusually large size county would tend to border on more counties than the average size county and would therefore be found in clusters more often than suggested by the tabled probabilities. In atypical situations such as these it might be reasonable to compare the obtained pattern with a large number of samples which included one or more specified counties in each sample selected. The resulting conditional probabilities would reflect this bias introduced by the inclusion of a constant county.

The computer program mentioned in the introduction allows for the inclusion of constant counties for obtaining these biased probability distributions.

Analysis of county contiguity within a state often shows clusters of counties along state boundaries. The question arises as to the results that might have been obtained had the bordering state been included in the analysis. The computer program allows for grouping a number of states into a single analysis and gives probability distributions of clusters and pairs of counties drawn from this larger parameter.

While geographic contiguity is not a common statistical problem, it is frequently encountered in epidemiology and geography. The "shaded map" seems to present a strong stimulus for "closure" or resolution, compelling the observer toward interpretation. In the experience of the authors, several observers may interpret a shaded map as showing various degrees of contagion. The most articulate, or, the senior observer present, wins the argument. The methodologies presented in this paper, when properly applied, provide the investigator with a simple, objective determination of contiguity.

REFERENCES

Geary, R. C. "The Contiguity Ratio and Statistical Mapping", Incorporated Statistician, Volume 5, 1954.

Ederer, Fred; Myers, Max H. and Mantel, Nathan. A Statistical Problem in Space and Time: Do Leukemia Cases Come in Clusters? 1963 (Unpublished).

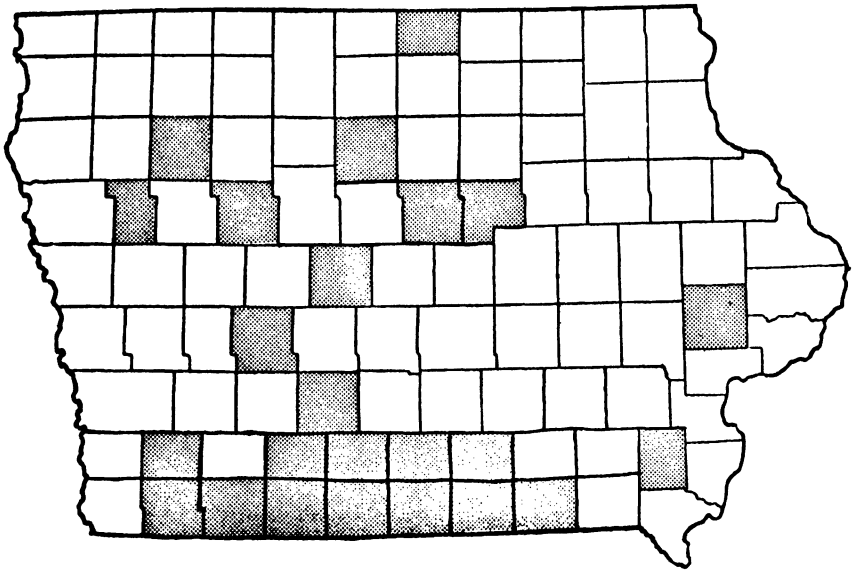
Table A

SAMPLE COUNTY CONTIGUITY LISTING

(A portion of the State of Iowa showing counties according to numeric code.)

INDEX COUNTY	COUNTIES CONTIGUOUS TO INDEX COUNTY					
01	15	39	61	88	02	
02	69	15	01	88	87	
03	96	22				
04	93	68	26			
05	83	14	39	15		
06	86	07	10	57	48	
07	38	12	09	10	06	86
08	37	94	40	85	77	25
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
99	46	41	40	35	94	

LIVE BIRTH RATE IN IOWA 1962
(lowest quartile)



Largest cluster - 13 counties

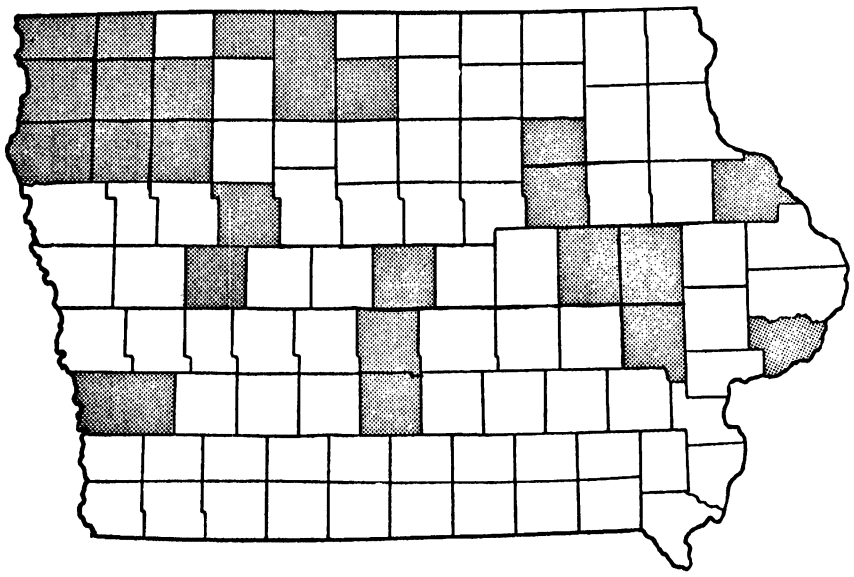
Probability that a cluster as large as 13 counties could have occurred by chance = .008

Number of pairs of counties - 17

Probability that as many as 17 pairs of counties could have occurred by chance = .117

Figure 1. A Configuration Favoring The Cluster Method

CRUDE DEATH RATE IN IOWA 1962
(lowest quartile)



Largest cluster - 8 counties

Probability that a cluster as large as 8 counties could have occurred by chance =.116

Number of pairs of counties - 20

Probability that as many as 20 pairs of counties could have occurred by chance =.018

Figure 2. A Configuration Favoring The Pairs Method

Table 1

ESTIMATED CUMULATIVE PROBABILITY DISTRIBUTION FOR CLUSTER SIZES OF CONTIGUOUS COUNTIES
SELECTED STATES - 25 PERCENT SAMPLE

(Based on 3,000 trials)

NUMBER OF COUNTIES IN LARGEST CLUSTER	ALABAMA[1] (17)	ARKANSAS (19)	CALIFORNIA (14)	COLORADO (16)	IOWA (25)	KENTUCKY (30)	LOUISIANA (16)	MINNESOTA (22)	MISSOURI (29)	MONTANA (14)	NEBRASKA (23)	PENNSYLVANIA (17)	SOUTH DAKOTA (17)	TENNESSEE (24)	VIRGINIA (24)	WISCONSIN (18)
1	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.005	.003	.037	.014	.001	.000	.146	.003	.000	.028	.002	.012	.014	.001	.003	.011
3	.103	.076	.255	.137	.062	.016	.164	.059	.018	.227	.080	.139	.142	.040	.073	.122
4	.319	.265	.543	.369	.254	.122	.426	.265	.134	.493	.306	.391	.397	.183	.270	.354
5	.538	.499	.767	.583	.492	.312	.646	.476	.327	.700	.536	.607	.607	.419	.497	.568
6	.701	.672	.879	.736	.688	.515	.793	.651	.511	.834	.725	.762	.769	.610	.687	.726
7	.820	.804	.947	.838	.806	.674	.913	.773	.672	.915	.826	.863	.866	.748	.814	.831
8	.896	.882	.976	.908	.884	.778	.937	.860	.785	.963	.898	.921	.923	.851	.882	.905
9	.946	.933	.991	.952	.931	.865	.966	.914	.860	.984	.938	.953	.960	.916	.928	.953
10	.974	.959	.997	.975	.952	.910	.988	.947	.912	.993	.966	.979	.979	.954	.957	.971
11	.988	.980	.999	.989	.977	.948	.994	.972	.947	.999	.982	.993	.990	.973	.978	.987
12	.995	.989		.995	.986	.968	.998	.985	.968		.986	.998	.995	.985	.986	.994
13	.998	.994		.998	.992	.982	.999	.993	.980		.995	.999	.999	.992	.990	.998
14	.999	.998		.999	.996	.991		.997	.989		.998			.995	.995	.999
15		.999			.998	.995		.998	.993		.999			.998	.998	
16					.999	.998		.999	.997					.999	.999	
17						.999			.999							

[1] Number in parentheses is the number of counties used in the 3,000 trials for that particular state and is approximately 25 percent of the total number of counties in the state.

Table 2

ESTIMATED CUMULATIVE PROBABILITY DISTRIBUTION FOR THE NUMBER OF PAIRS OF CONTIGUOUS COUNTIES
SELECTED STATES - 25 PERCENT SAMPLE

(Based on 3,000 trials)

NUMBER OF PAIRS	ALABAMA[1] (17)	ARKANSAS (19)	CALI-FORNIA (14)	COLORADO (16)	IOWA (25)	KENTUCKY (30)	LOUI-SIANA (16)	MINNE-SOTA (22)	MISSOURI (29)	MONTANA (14)	NEBRASKA (23)	PENN-SYLVANIA (17)	SOUTH DAKOTA (17)	TENNESSEE (24)	VIRGINIA (24)	WISCONSIN (18)
1	.000	.000	.001	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000
2	.000	.000	.004	.001	.000	.000	.001	.000	.000	.003	.000	.001	.000	.000	.000	.000
3	.001	.000	.028	.003	.000	.000	.004	.001	.000	.018	.000	.003	.001	.000	.000	.002
4	.004	.001	.083	.015	.000	.000	.013	.001	.000	.064	.000	.012	.009	.000	.000	.009
5	.016	.004	.196	.048	.001	.000	.053	.005	.000	.155	.002	.032	.033	.000	.001	.023
6	.052	.016	.353	.104	.004	.000	.137	.011	.000	.298	.010	.080	.089	.001	.003	.061
7	.122	.050	.523	.216	.012	.000	.260	.031	.000	.470	.027	.170	.191	.004	.007	.128
8	.234	.109	.685	.361	.031	.000	.415	.064	.001	.629	.069	.310	.320	.012	.017	.245
9	.368	.214	.815	.517	.074	.002	.566	.139	.005	.758	.147	.464	.471	.031	.045	.371
10	.523	.344	.894	.657	.144	.004	.721	.231	.009	.860	.245	.615	.625	.073	.094	.528
11	.664	.495	.948	.778	.238	.012	.822	.338	.024	.925	.359	.745	.754	.137	.164	.674
12	.779	.637	.976	.863	.354	.031	.896	.478	.052	.964	.501	.842	.852	.227	.272	.786
13	.868	.746	.991	.926	.493	.060	.941	.597	.096	.982	.631	.912	.923	.345	.404	.871
14	.930	.835	.996	.962	.632	.111	.967	.715	.161	.993	.756	.952	.958	.467	.521	.923
15	.967	.901	.998	.982	.752	.183	.983	.805	.256	.997	.851	.977	.978	.592	.641	.956
16	.982	.943	.999	.990	.842	.277	.991	.884	.356	.999	.906	.990	.988	.703	.748	.978
17	.992	.970		.997	.883	.382	.996	.922	.480		.943	.994	.996	.804	.832	.987
18	.996	.986		.999	.938	.508	.998	.954	.587		.969	.996	.998	.868	.897	.995
19	.998	.994			.965	.617	.999	.974	.690		.984	.999	.999	.921	.935	.999
20	.999	.998			.982	.719		.985	.770		.993			.949	.964	
21		.999			.993	.798		.993	.846		.997			.972	.980	
22					.996	.861		.995	.899		.998			.985	.991	
23					.998	.909		.999	.935		.999			.994	.995	
24					.999	.942			.961					.997	.997	
25						.943			.976					.997	.999	
26						.983			.987					.999		
27						.989			.991							
28						.995			.996							
29						.998			.998							
30						.999			.999							

[1] Number in parentheses is the number of counties used in the 3,000 trials for that particular state and is approximately 25 percent of the total number of counties in the state.