This research was supported by the National Institutes of Health, through the Division of Environmental Health Sciences, Grant No. 329-NIH-300-3.

A COMPUTER APPLICATION OF KNOX'S TEST OF CLUSTERING

Antonie Wouter Voors
University of North Carolina

Institute of Statistics Mimeo Series No. 564

January 1968

A COMPUTER APPLICATION OF KNOX'S TEST OF CLUSTERING [1]

A. Wouter Voors

University of North Carolina

The finding of time-space clustering of a certain disease in a population of individuals, in the absence of simple time clustering and simple space clustering, restricts the possible explanations of this phenomenon to only two: (a) one diseased individual causes the disease in the next one, and (b) the disease is caused by some outside determinant not brought about by diseased individuals and only operating in this population at limited times and places. This very restriction could mean a step forward in the search for either the etiology or the control of this disease. Similar reasoning can be applied when "time" is replaced by some other index of distance, using a short time period of observation. A method of testing for such clustering, using data collected routinely, is described.

1. DESCRIPTION OF THE PROBLEM

Consider a population of individuals. Let each individual have four characteristics: one dyadic characteristic indicating whether or not the individual has a certain disease, one temporal characteristic (time of observation), and two spatial characteristics being the two coordinates of a rectangular coordinate system. Now consider the joint distribution of individuals over the four characteristics. Here it is of interest to assess whether there is time-space clustering of diseased individuals in this distribution. If there were such clustering, this would mean that the presence of one diseased individual enhances the probabilty of other cases to occur nearby soon. Such a finding would allow one of only two

---

[1] I am thankful to Dr. Dana Quade, Associate Professor of Biostatistics School of Public Health, University of North Carolina, for his direction in finding the above described method, and to Mr. G. Lindsay Cleveland, Jr. for his advice in computer programming.

possible explanations: (a) one diseased individual causes the disease in the next one, or (b) the disease is caused by some outside determinant not brought about by diseased individuals and operating only at limited times and places.

Similar reasoning can be applied when "time" is replaced by some other index of distance between two cases, while the time period of observation is small, that is to say, not much larger than the suspected average duration of the infectivity in any one case, or of the operation of any other postulated cause.

2. LITERATURE REVIEW

Ederer, Myers and Mantel [1] observed a number of towns over a number of years for cases of a certain disease. They assumed implicitly that all towns had an equal and fixed number of persons at risk. They concluded non-clustering when the occupancy of cases per town-year was not significantly different from the expectation on a random basis.

Barton and David [2] considered an orthogonal system of two spatial and one time dimension. They subdivided the time axis into n equal intervals j where $j = 1, 2, \ldots n$. Let each interval have c_j disease cases i, where $i = 1, 2, \ldots c_j$, each case having a distance d_i from the point with over-all mean space coordinates. They considered the distribution of

$$Q = \frac{N - 1}{N - n} \left[1 - \frac{\sum_{j=1}^{n} \left\{ \frac{1}{c_j} \left(\sum_{i=1}^{c_j} d_i \right)^2 \right\}}{\sum_{j=i}^{n} \sum_{i=1}^{c_j} d_i^2} \right]$$

obtained by taking all possible ways in which the $N = \sum_{j=1}^{n} c_j$ cases can be assigned to the n groups, and thus they designed a cluster test. I believe that this test is insensitive to a clustering near the point of

mean space coordinates.

Knox [3], [4] described the following test for clustering. All possible combinations of the observed cases, taken two at a time, are entered into a 2 x 2 table by interval between times of onset and by distance between residences. He applied to this table a Poisson test, based on the assumption that the smallest expected frequency in the table cell of smallest time interval and smallest distance follows the Poisson distribution. The validity of the Poisson assumption, despite lack of independence of the paired case parameters, was confirmed empirically by Pike and theoretically by David and Barton [5]. It would appear that the assumption holds for small expected frequencies only.

All of the above-mentioned references do not take into account a possibly uneven distribution of persons-at-risk over the units of space. In practice, this implicit assumption of homogeneous distribution is likely to make these tests less sensitive (less powerful): when, say, exposure to rural flora or fauna would cause rural case clustering, the implicit assumption would hinder the detection of this clustering. When, on the other hand, central urban slum living would cause urban case clustering, the same tests would be pronounced invalid due to violation of the assumption.

3. METHOD OF PROCEDURE

Consider all possible combinations of the diseased cases taken two at a time and tabulate the pairs in a contingency table (the "observed" table) by difference in determinant 1 (say, time) and by difference (distance) in determinant 2 (say, space). Now the number of pairs in the table cell of small time difference and small space distance is compared with the number expected under the hypothesis of independence between the two effects.

A frequency distribution of the expected values in this "cell of close pairs" is generated as follows. Consider the marginal values in a contingency table of observed cases by determinant No. 1 (time) and determinant No. 2 (space). These observed values are used for estimating the time-specific and place-specific rates of occurrence. Thus the marginal values in any corresponding "expected" table will be identical to those in the "observed" table. Now the question is how the "expected" cases are distributed over the non-marginal cells. The answer is that each possible way of pairing determinant No. 1 with determinant No. 2 should have equal probability in arriving at the expected case distribution over the two determinants, under compliance with the constraint of the given marginal values. Thus, if n cases are observed, and if their observed joint distribution is represented in the contingency table of "time" rows and "space" columns, then the total of all marginal values both row and column-wise is n and there are n! possible ways of pairing any time location with any space location such that all marginal table values are preserved.

4. EXAMPLE 1: ARTIFICIAL PROBLEM

An elementary example of this "expected" distribution is as follows. Suppose that 4 cases of a certain disease occurred in an area covering 4 census tracts, namely 2 cases in census tract No. 1, 1 case in census tract No. 2, no cases in census tract No. 3, and 1 case in census tract No. 4. Suppose further that these 4 cases were observed over a period of 4 months, namely 1 case in month No. 1, no case in month No. 2, 2 cases in month No. 3, and 1 case in month No. 4. The distribution is given in Table 1.

The various ways in which the months and census tracts can be paired in the 4 cases under maintenance of the same marginal distribution is given in Figure 1. The total number of possible ways of pairing is 24=4! Define that two cases are proximate in time if, and only if, they occur in equal or adjacent months; and that two cases are proximate in space if, and only

TABLE 1

OBSERVED CASES BY MONTH AND CENSUS TRACT *

Census Tract No.	Month No.				
	1	2	3	4	total
1	0	1	0	0	1
1	0	0	1	0	1
2	0	0	0	1	1
4	1	0	0	0	1
total	1	1	1	1	4

*Fictitious data for illustrative purpose.

FIGURE 1

The 24 Different Ways in Which the Months and Census Tracts of Table 1 Can Be Combined

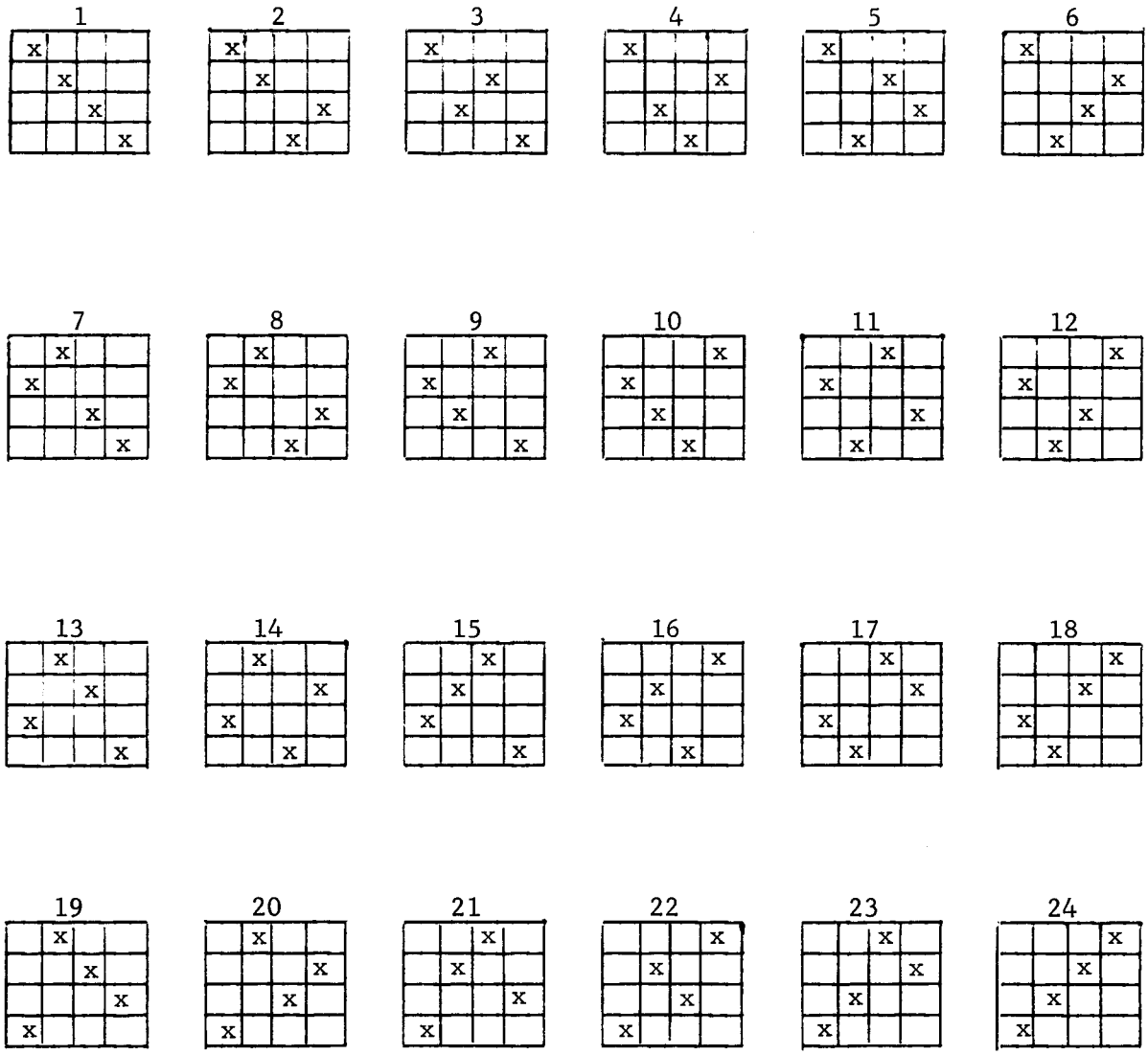


TABLE 2

"PROXIMATE PAIRS" FOR THE 24 POSSIBLE WAYS IN WHICH FOUR GIVEN CENSUS TRACTS CAN BE COMBINED WITH FOUR MONTHS*

Combination No.**	Proximate Pairs***
1	1
2	0
3	1
4	0
5	1
6	1
7	1
8	0
9	1
10	0
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	3
20	3
21	3
22	3
23	3
24	3
total	32

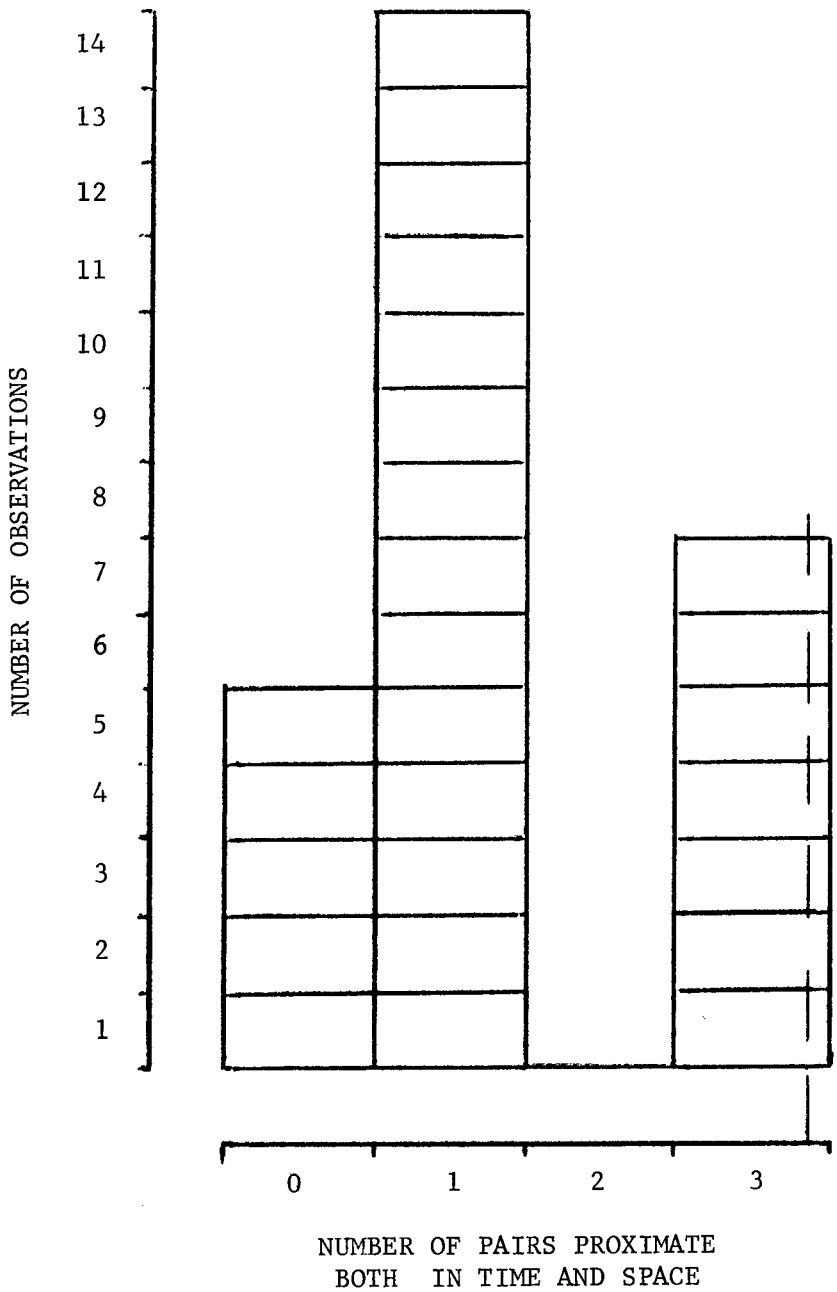
* Data from table 1.

** The numbers by which the various combinations are indicated correspond to those of figure 1.

*** For definition of proximity see text.

FIGURE 2

Frequency Histogram of Possible Observations on Number of Pairs Proximate Both in Time and Space. *



* Data from table 1.
The interrupted line delineates the right-tail 5% of the histogram surface area.

if, they occur in census tracts bearing numbers differing one or less. Table 2 lists the number of case pairs proximate both in time and in space for the 24 possible ways of pairing. From this list a frequency distribution of such proximate pairs can be constructed (Figure 2). In the present example, 3 of such proximate pairs would be expected to occur 6 times out of 24 observations if the population at risk is equally distributed over the census tracts. Thus, the statistical significance level of 5% is not reached.

5. METHOD OF PROCEDURE --- CONTINUED

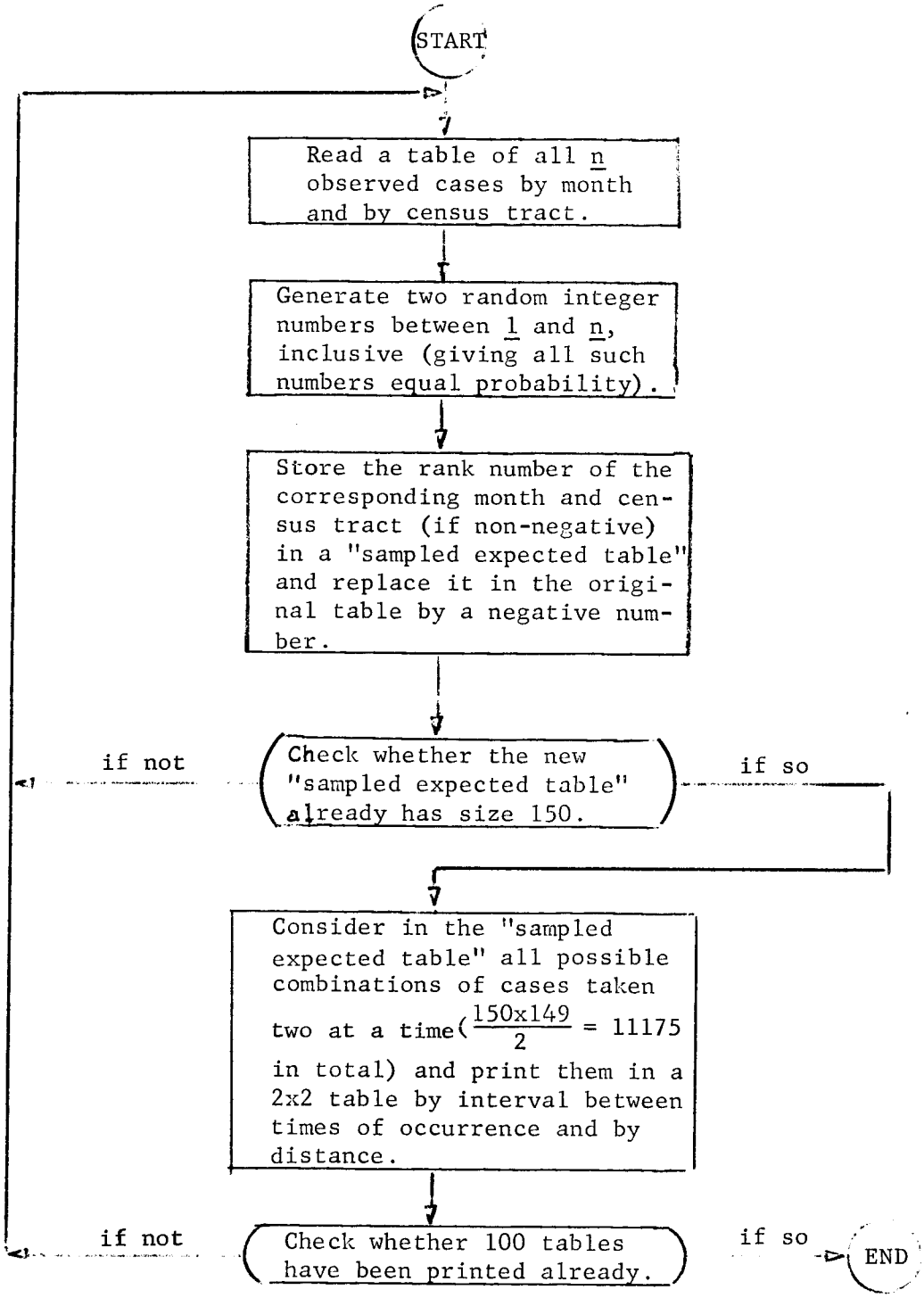
When there are n observed cases, they can form $\binom{n}{2} = \frac{n(n-1)}{2}$ possible combinations of cases taken two at a time. A convenient way of measuring space distances between the residences of such case pairs is the use of an electronic computer. If for each residence or for each center of its census tract the two rectangular coordinates are known, distances between pairs of residences or centers can be calculated by using the Theorem of Pythagoras. Once such a computer has been put to work, it is relatively easy to take a random sample from the total of possible pairings of time units (months) and census tracts.

A flow chart of the relevant computer program is given in Figure 3.

The result can be presented in a frequency histogram of a form similar to the one in Figure 2. The location on the horizontal beyond with only 5% of the observations occurred can be calculated (interrupted line in Figure 2). The sampling error of this location can be estimated by repeating the whole procedure several times, and by estimating from

FIGURE 3

Flow Chart of a Computer Program Generating 100 Random Samples, Each Consisting of 150 "Cases" That Form 11175 Pairwise Combinations of Which an "Expected" Number are Proximate Pairs. *



* It is assumed that the number n of cases is larger than 150. A random sample consisting of 150 cases is taken from each one complete pairing consisting of n month-census-tract pairs.

these samples the standard error of the mean location.

6. EXAMPLE 2: AGE-SPACE CLUSTERING OF IN-SITU CERVICAL CARCINOMA

Description of the Problem

A necessary condition for the direct transmission of infectious diseases is the presence of the infective and the infected person at the same place at the same time. Conversely, a disease of unknown causation can be examined for contagiousness by assessing whether or not the place and time of a new patient at the onset of his incubation period coincided with the place and time of an old patient in his infective stage. In such studies address of residence is often used as an index of the patient's location (Knox [3], [4], David and Barton [5], Lundin et al [6]). Moreover, difference in age can function as an additional index of social distance between patients (Ipsen [7]).

If a disease of unknown origin is suspected to have a causal factor which is venereally transmitted (such as cervical carcinoma), the assumptions underlying the use of age proximity as an index of socio-sexual proximity postulates that two females of proximate age are more likely to have a common male partner than females of non-proximate age.

In a previous report [9] attention has been drawn to the age-space clustering of cervical carcinoma cases in Rochester, New York as reported for the years 1962-1964. It was argued that this relationship is consistent with and suggestive of a venereally transmissible factor in the causation of cervical carcinoma. In the present investigation the number of observed cases has been expanded (see below).

Materials and Methods--Observation

The cervical cancer cases in the city of Rochester, New York, reported for the years 1942-1965 are given by triennial periods by age and by clinical stage (in-situ versus invasive) in Table 3. The present study consists of all in-situ cases reported in years with a reasonably high proportion of in-situ cases, namely the years 1957-1965.

For each case in the study the following data were punched on cards: a) age of reporting; b) census tract of residence.

The computer performed the following operations:

(A) The patient's place of residence was approximated by substituting the geometrical center of the census tract in which the patient resided. If this approximation entails bias, this bias is on the conservative side [9]. The census tract center was represented by its location in a system of North-South and East-West coordinates, each 0.1 mile apart.

(B) Each possible combination of two cases was considered, and for each the following two items were recorded: a) the difference in age of reporting, and b) the distance in miles between the geometrical centers of the respective census tracts.

(C) The combinations were tabulated according to the two items mentioned.

Material and Methods--Expectation

One hundred and fifty artificial cases were generated by a random choice of age of reporting (giving each reported individual age an equal

TABLE 3

REPORTED CERVICAL CARCINOMA CASES BY YEAR, STAGE AND AGE OF REPORTING, CITY OF ROCHESTER, NEW YORK, 1942-1965*

Year of reporting	Stage	Age at reporting (in years)							
		15-24	25-34	35-44	45-54	55-64	65-74	75+	total
1942-'44	in-situ	0	0	0	0	0	0	0	0
	invasive	0	7	29	31	41	16	9	133
1945-'47	in-situ	0	0	0	0	0	0	0	0
	invasive	1	3	32	40	31	23	8	138
1948-'50	in-situ	0	1	1	0	0	0	0	2
	invasive	0	5	27	41	38	21	13	145
1951-'53	in-situ	0	1	2	1	1	0	0	5
	invasive	0	3	22	39	38	17	10	129
1954-'56	in-situ	0	6	17	2	1	2	0	28
	invasive	0	11	15	33	22	24	20	125
1957-'59	in-situ	2	25	30	11	7	4	1	80
	invasive	0	13	17	23	22	22	7	104
1960-'62	in-situ	2	21	32	14	4	0	1	74
	invasive	0	2	20	21	24	18	7	92
1963-'65	in-situ	0	29	30	15	17	2	1	94
	invasive	0	8	23	16	21	18	8	94
total	in-situ	4	83	112	43	30	8	3	283
	invasive	1	52	185	244	237	159	82	960

* New York State Department of Health

chance to be included) and likewise an independently random choice of residence (Figure 3).

The double random choices and subsequent combinatorial and tabulatorial operations were performed 100 times, and the whole procedure was repeated five times.

The five resulting frequency distributions of entry into the cell of space- and time-adjacent pairs are presented in Figure 4.

Results

There were 248 cases of in-situ cervical carcinoma reported in Rochester during the years 1957 through 1965. Hence there were $\binom{248}{2}$ = 30,628 possible combinations, taken two at a time. These possible combinations of observed (reported) cases are given by index of distance between residences and by difference of age of onset (Table 4).

It is apparent in the cells of Table 4 that 614 out of 30,628 pairs are observed to be adjacent both in age and space, which makes up 2.0047 % of all pairs, or 224.025 out of each set of 11175 pairs are proximate in age and time of onset.

The five frequency distributions of possible pairs of cases proximate as to residence and age, expected for Monroe County under assumption of space-time independence, are given in Figure 4. Here, the 5% points are respectively at the following numbers of age-space proximate pairs: 226.7, 225.0, 221.7, 235.0 and 222.5. Assuming normal distribution (which assumption is justifiable according to the central limit theorem), the estimated confidence interval of the mean expected 5% point at the 5% level of significance is $226.50 \pm (1.96)\ (5.403) =$ (215.9, 237.1). Since the observed point is 224.0 (see above), the

FIGURE 4

Five Frequency Distributions of 100 "Expected" Numbers of Case Pairs with Proximate Ages and Residences, Generated Independently on a Random Basis under the Assumption of Independence between Age and Space, while Adhering to the "Observed" Marginal Values. In-Situ Cervical Carcinoma Reported in Rochester, New York, 1957-65. Each of these "Expected" Numbers of Age- and Space-Proximate Pairs Applies to a Total of 11175 Combinations of Two Generated Cases.

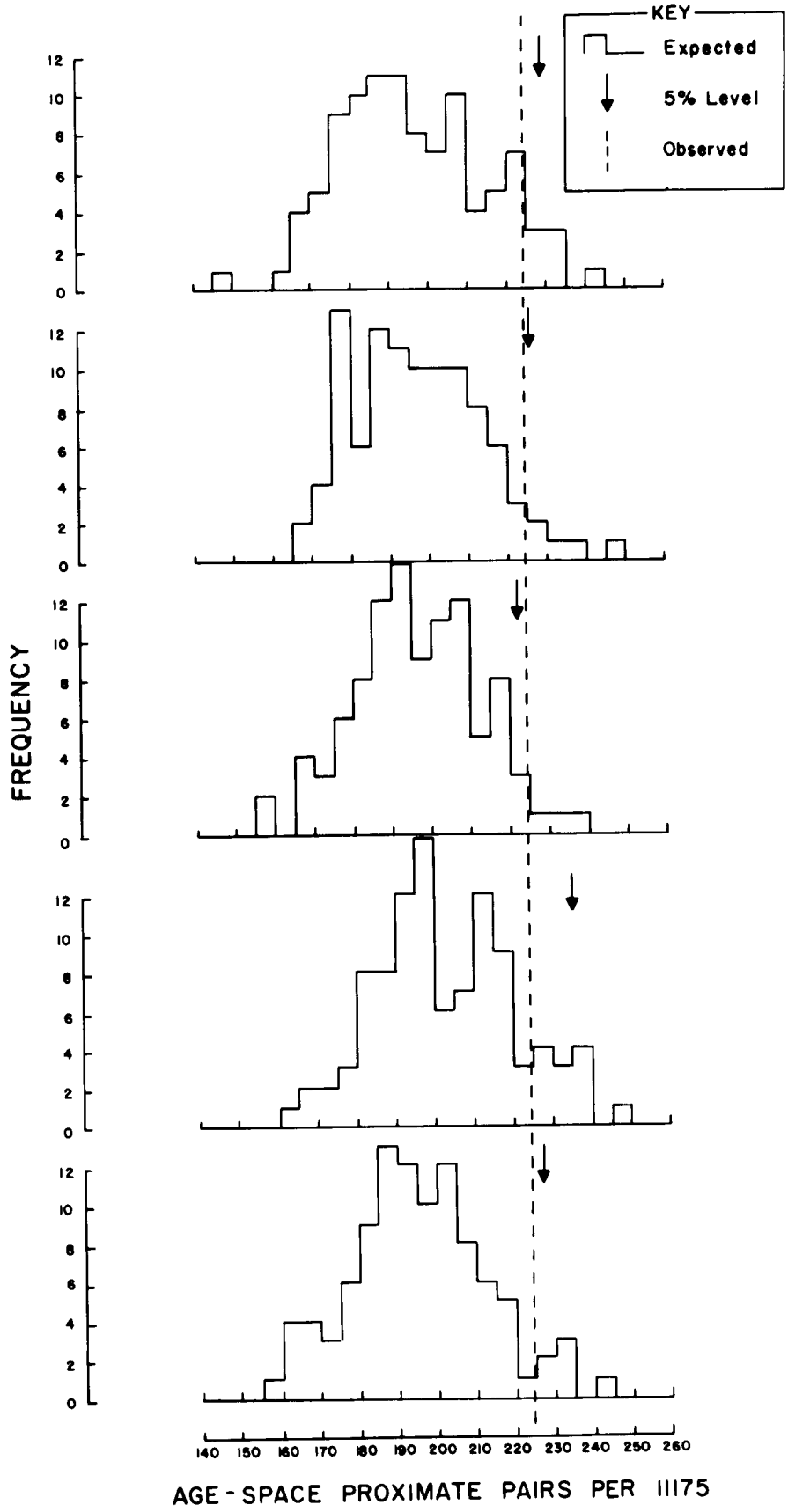


TABLE 4

PAIRS OF REPORTED IN-SITU CERVICAL CANCER BY AGE DIFFERENCE AND DISTANCE ROCHESTER, NEW YORK, 1957-1965.

Age of earlier reported case minus age of later reported case (in years)	Distance between geometrical centres of census tracts		
	<1 mile	>1 mile	total
33+	105	707	812
32-- 28	98	529	627
27-- 23	108	784	892
22-- 18	207	1187	1394
17-- 13	316	1910	2226
12-- 8	413	2587	3000
7-- 3	505	3293	3798
2-- -2	614	3458	4072
-3-- -7	507	3304	3811
-8-- -12	410	2805	3215
-13-- -17	305	2157	2462
-18-- -22	198	1513	1711
-23-- -27	129	951	1080
-28-- -32	82	651	733
-33+	79	716	795
Total	4,076	26,552	30,628

observed clustering is not significant at the 5% level.

If Figure 4 represents true probability distributions, the observation is consistent with acceptance of the "non-cluster" hypothesis at the 5% level of significance.

Conclusion

It could not be disproved that the excess of in-situ cervical carcinoma cases, adjacent as to age and residence, as observed for Rochester, New York, for the years 1957-1965, is due to chance alone.

7. SIGNIFICANCE OF THIS RESEARCH

The increasing extent to which researchers have access to electronic computer service makes feasible the sensitive testing of certain hypotheses concerning the clustering of individuals when distributed jointly according to two characteristics known to be determinants of the state (for instance, the disease) in which these individuals are. This testing is a relatively straight-forward procedure requiring data often available from ongoing routine collection of data.

In the health field, the gained insight in disease causation as derived from the testing for clustering may be of interest either from the viewpoint of etiologic theory or of disease control. For instance, the question whether there is such clustering by time of occurrence and place of residence in leukemia cases would elucidate the etiology of this disease [1]. On the other hand, the question whether communicable respiratory disease plays a role in the causation of cardiovascular death might influence the current policies of both infectious and cardiovascular disease control, even without full under-

standing of the underlying mechanism. The existance of such a role would be supported -- although not proven -- by the eventual demonstration of time-space clustering of the cardiovascular deaths.

REFERENCES

[1] Ederer, F.; Myers, M. H. and Mantel, N.: A statistical problem in space and time: do leukemia cases come in clusters? Biometrics, 20: 626-638 (1964).

[2] Barton, D.E.; David, F.N. and Merrington, M.: A criterion for testing contagion in time and space. Ann. Hum. Genet. Lond., 29: 97-102 (1965).

[3] Knox, G.: Epidemiology of childhood leukemia in Northumberland and Durham. Brit. J. Prev. Soc. Med., 18: 17-24 (1964).

[4] Knox, G.: Detection of low-intensity epidemicity. Brit. J. Prev. Soc. Med., 17: 121-127 (1963)

[5] David, F.N. and Barton, D.E.: Two space-time interaction tests for epidemicity. Brit. Jour. Prev. Soc. Med., 20: 44-48 (1966).

[6] Lundin, F.E. Jr., Fraumeni, J.F. Jr., Lloyd, J.W., and Smith, E.M.: Temporal relationships of leukemia and lymphoma deaths in neighborhoods. J. Nat. Cancer Inst., 37: 123-133, (1966).

[7] Ipsen, J. Jr.: Social distance in epidemiology. Human Biol., 31: 162-179, (1959).

[8] Kinsey, A.C., Pomeroy, W.B., and Martin, C.E.: Sexual behavior in the human male. Philadelphia, Saunders, (1948).

[9] Voors, A.W.: Interaction between geographic and age proximity in cervical carcinoma. Am. J. Epid., 85: 387-394, (1967).