Evaluating Coinciding Anomalies Along a Fault Trace or Other Traverse—Simulations and Statistical Procedures

U.S. GEOLOGICAL SURVEY BULLETIN 1802

U.S. DEPARTMENT OF THE INTERIOR
March 3, 1849

Evaluating Coinciding Anomalies Along a Fault Trace or Other Traverse—Simulations and Statistical Procedures

By RUSSELL L. WHEELER and KATHERINE B. KRYSTINIK

Methods for evaluating spatial associations of anomalies in several different types of data along a linear or curvilinear traverse to determine whether the anomalies coincide by chance

U.S. GEOLOGICAL SURVEY BULLETIN 1802

DEPARTMENT OF THE INTERIOR
DONALD PAUL HODEL, Secretary

U.S. GEOLOGICAL SURVEY
Dallas L. Peck, Director

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON: 1987

For sale by the
Books and Open-File Reports Section
U.S. Geological Survey
Federal Center
Box 25425
Denver, CO 80225

Library of Congress Cataloging-in-Publication Data

Wheeler, Russell L.
Evaluating Coinciding Anomalies Along a Fault Trace or Other Traverse—Simulations and Statistical Procedures.
(U.S. Geological Survey Bulletin ; 1802)
"Methods for evaluating spatial associations of anomalies in several different types of data along a linear or curvilinear traverse to determine whether the anomalies coincide by chance."
Bibliography: p.
Supt. of Docs. no.: I 19.3:1802
1. Geology—Statistical methods. 2. Spatial analysis (Statistics) I. Krystinik, Katherine B. II. Title. III. Series.
QE75.B9 no. 1802 557.3 s 87-600288
[QE33.2.M3] [550'.72]

CONTENTS

FIGURES

TABLES

Evaluating Coinciding Anomalies Along a Fault Trace or Other Traverse—Simulations and Statistical Procedures

By Russell L. Wheeler *and* Katherine B. Krystinik

Abstract

If two or more types of data are mapped along the same traverse, such as a fault trace, anomalies in different data types can coincide. If the anomalies coincide because they have a common geological cause, the spatial coincidences are worth interpreting to understand that cause. Alternatively, the anomalies might coincide by chance, so that genetic interpretations of the observed patterns of anomalies would be misguided. Opinion is an unreliable guide for deciding which is the case.

Numerical simulations and statistical tests aid in this decision. Each simulation produces a random scattering of the observed anomalies along the traverse and an anomaly pattern that resembles, to some degree, the one observed. Randomization tests are used to compare the observed pattern of anomalies to the collection of simulated patterns, in terms of numbers of triplets and larger groups of coincident anomalies. Test results show whether the anomalies of the observed pattern coincide more than should be attributed to chance. If so, the group of coincident anomalies is worth interpreting.

Jaccard coefficients detect those data types that tend to have anomalies together along the traverse. Two such highly associated data types can be combined. Together they might detect the occurrence of whatever causes the coincident anomalies more effectively than can either data type alone.

INTRODUCTION

A problem that is commonly encountered in the earth sciences is the evaluation of spatial associations between features on maps of two or more kinds of data that are collected over the same study area. For features that are represented on the maps as points, geographers and statisticians have developed many procedures for detecting and evaluating spatial associations (for example, Lewis, 1977; Ripley, 1981). The problem becomes successively more complex if the mapped features are represented as lines, as patches of similar sizes and simple, similar shapes, or as irregular areas of dissimilar sizes and shapes. Figure 1 illustrates this last case, which is a common one in geology. Anomalies in all three data types appear to coincide in the upper left part of the study area. Pairs of anomalies appear to coincide at two places on traverse *A-A*′. Are these spatial coincidences worth interpreting, or should they be dismissed as chance overlaps of unrelated anomalies in different types of data?

The general problem has four parts: (1) defining anomalies, (2) identifying coincident anomalies, (3) determining whether the coincident anomalies are likely to have coincided by chance instead of from some common cause, and (4) interpreting such coincidences. Solving each part requires solving all the preceding parts. The last part is usually the most fun, which might explain the common tendency to jump to it without explaining how (or if) the first three parts of the problem were solved. The resulting arguments are often more exciting than informative, perhaps because they might root in unrecognized disagreements about the first three parts of the problem.

The evaluation of earthquake hazards encounters a simplified version of this problem. Instead of a map area, we have the trace of a fault on the Earth's surface, for example traverse *A-A*′ in figure 1. Data were

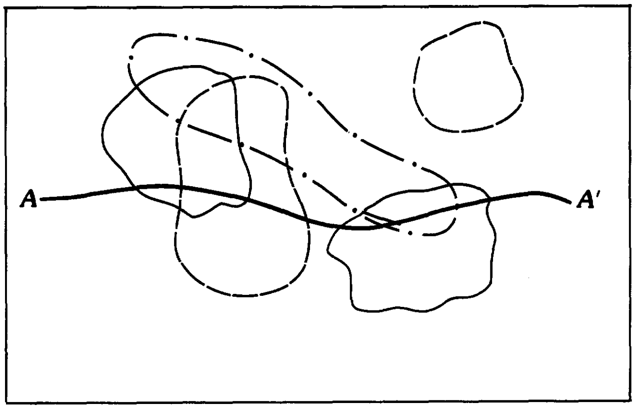

Figure 1. Map made by superimposing three hypothetical maps of the same study area. Each superimposed map shows spatial distribution of a different type of geological, geophysical, or other earth-science data. Each data type has anomalies in one or more parts of study area. Solid, dashed, and dashed-dotted lines outline anomalies in various data types, one line type per superimposed map. *A-A*′ is traverse discussed in text.

collected or measured along the trace, or within a narrow strip around it. The problem and its four parts are unchanged.

For anomalies that might coincide along a study traverse like A-A' in figure 1, we present a general procedure for solving the second and third parts of the problem. Most such traverses will be along fault traces, but the procedure can be used along any kind of linear or curvilinear traverse.

We will not treat the first part of the problem, that of defining the anomalies, because the methods for doing that depend on the types of data to be analyzed, on the ways in which the data are represented or summarized on maps, and on the goals of the analysis. For example, this report emerged from an attempt to identify places along the Wasatch fault zone in Utah where the fault zone might be segmented into lengths that tend to rupture independently of each other (Schwartz and Coppersmith, 1984; Wheeler, 1984; Machette and others, 1986). The goal was to detect boundaries between segments. The approach was to examine kinds of data that ought to record the presence of a segment boundary as anomalies of predictable sorts. This goal imposed conditions on the kinds of data to be analyzed and on the kinds of anomalies to be sought in each data type. Wheeler and Krystinik (1987a, b, in press) used anomalies along the north-striking fault zone that include east-trending gravity and aeromagnetic gradients and large cross faults, and changes along the fault zone in abundance of earthquake epicenters, in the geometry of the fault zone, and in the height and width of the upthrown fault block. Because many anomalies were identified subjectively by inspection of maps of points, contour maps, and geologic maps, the anomalies had to be shown to be reliable, for example by being recognized by two or more independent and competent workers. Wheeler and Krystinik (in press) described how the data and anomalies from the Wasatch fault zone satisfied these conditions. Peculiarities of other investigations will impose their own conditions on data and anomalies.

The second part of the problem, that of identifying coincident anomalies, requires specifying objective rules for identification. We did this by developing a procedure for identification of coincident anomalies, programming the procedure, and testing the program on artificial data.

The third part of the problem, determining whether the observed coincidences of anomalies are or are not likely to have occurred by chance, is a statistical effort. If anomalies are randomly scattered, the exact form of the distribution that underlies the numbers of coincident anomalies is unknown, so we solve this part of the problem with simulations and statistical tests of the simulated anomaly patterns. (We describe the simulation and statistical procedures in detail because this report is directed toward geophysicists and quantitatively oriented geologists, many of whom have little training in statistics.) The statistical nature of this part of the problem imposes two more conditions on the anomalies (Wheeler and Krystinik, in press). Some version of these two conditions will apply to any investigation of any types of data that can be represented as in figure 1. The anomalies in any one data type had to be identified without reference to any other data type used in the analysis, and also without reference to the segment boundaries that Schwartz and Coppersmith (1984) suggested in the report that prompted our investigation.

The fourth part of the problem is that of interpreting the group of coincident anomalies that is unlikely to be an artifact of chance. For the Wasatch fault zone, Wheeler and Krystinik (1987a, b) applied the methods of this report to the data of Wheeler and Krystinik (in press) and interpreted the results.

COINCIDENT ANOMALIES

For each data type, the presence and absence of anomalies along the length of the fault trace or other study traverse are summarized graphically. In an example (fig. 2), three anomalies of different widths are located where variable v(i) has a value of 1. Along the other four sections of the traverse a value of 0 for v(i) records the absence of any anomaly in this data type.

We need a rule to determine when anomalies coincide. We define two anomalies as coincident if either anomaly contains the center of the other (fig. 3). For example, the two anomalies labelled 1 in figure 3 coincide by this rule, because the wide anomaly in v(2) includes the center of the narrower anomaly in v(1). This example illustrates why the rule does not require each anomaly to contain the center of the other. The narrow anomaly in v(1) does not contain the center of the wide anomaly in v(2), but these anomalies coincide in any geologically reasonable sense of the word. The second two anomalies coincide because each contains the center of the other. The third two anomalies do not coincide because neither contains the center of the other. Finally, two anomalies in the same variable cannot coincide because they cannot merge or overlap without becoming a single anomaly.

We use a hypothetical pattern of anomalies in four data types (fig. 4) to illustrate procedures throughout this report. Each data type contains from one to six anomalies, which vary in width and separation. The 17 anomalies have a median width of 3 km, and range from 1 to 6 km in width except the single 30-km wide anomaly (K) in data type 3 (table 2). Depending on the data type, from 70 to 83 percent of the traverse's length lacks anomalies. Pairs of coincident anomalies are found at four places along the traverse (fig. 4). At about d= 32 three types

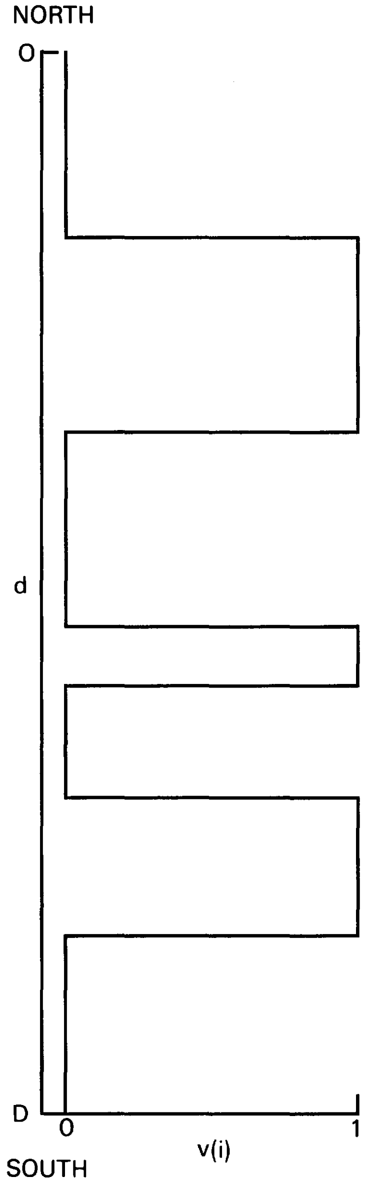

Figure 2. Sketch showing locations of several anomalies in one data type (such as along a fault trace or other traverse of a study area). Table 1 defines symbols.

of data appear to be anomalous, and at about d= 55 all four types appear to be anomalous.

A triplet of coincident anomalies can be identified by extending the rule for coincident pairs in this manner: if two anomalies in different variables coincide, a third coincides with both if it coincides with each. The extended rule has several consequences that will form the bases of later analyses. For the three anomalies at d= 32 (fig. 4), there are three possible pairs of coincident anomalies, and all three coincident pairs exist. The anomaly (C) in v(1) spans km 30-35, (H) in v(2) spans

Table 1. Symbols used in representation of anomalies

Symbol	Definition
n	Number of data types to be examined for anomalies.
d	Distance along traverse in km, from d = 0 km at north end of traverse to d = D km at south end.
v(i)	Binary variable that can take values of 0 or 1 along traverse. In lengths of traverse that cross an anomaly in data type i, where i=1, ..., n, v(i)=1. In lengths between anomalies in data type i, v(i)=0. Where v(i)=1, v(i) is said to have an anomaly or to be anomalous.

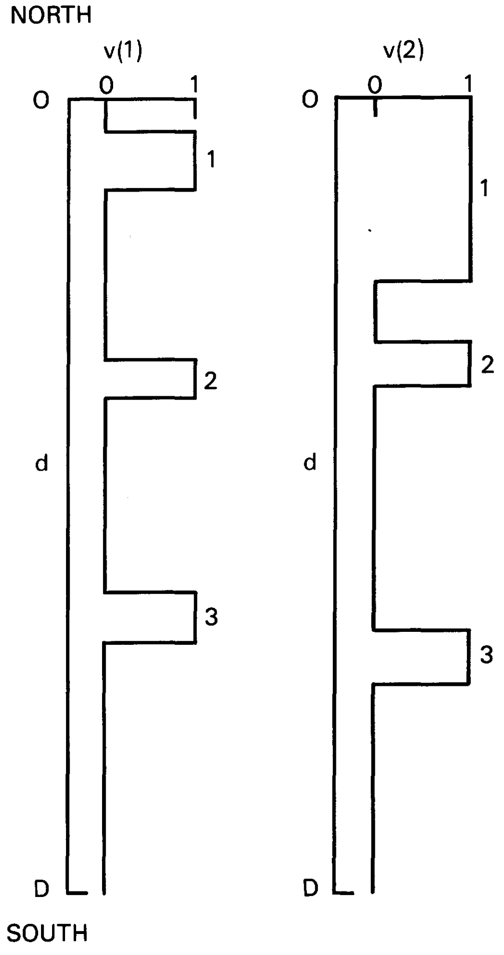

Figure 3. Sketch illustrating identification of coincident anomalies. Table 1 defines symbols. Two types of data occur along traverse, and anomalies in them are located by values of v(1) and v(2). Numerals identify pairs of anomalies that fall at about the same places along traverse.

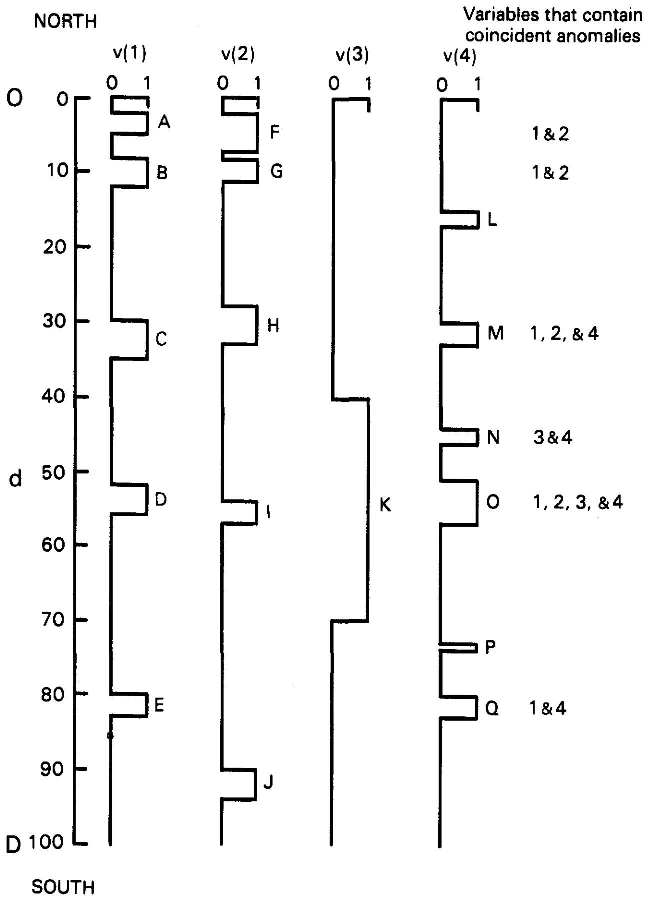

Figure 4. Hypothetical anomaly pattern in four data types along a traverse. Table 1 defines symbols. Data of types 1-4 have 17 anomalies along length of traverse. Anomalies are where variables v(1) to v(4) have values of 1, by the convention illustrated in figure 2. Letter to right of each anomaly keys it to table 2. Column at right shows where and which anomalies coincide.

Table 2. Descriptions of anomalies illustrated in figure 4

Anomaly	Distance from 0 (km) Ends	Center	Width (km)	Coincident with
		v(1)		
A	2-5	3.5	3	F
B	8-12	10	4	G
C	30-35	32.5	5	H,M
D	52-56	54	4	I,K,O
E	80-83	81.5	3	Q
		v(2)		
F	2-7	4.5	5	A
G	8-11	9.5	3	B
H	28-33	30.5	5	C,M
I	54-57	55.5	3	D,K,O
J	90-94	92	4	-
		v(3)		
K	40-70	55	30	N,D,I,O
		v(4)		
L	15-17	16	2	-
M	30-33	31.5	3	C,H
N	44-46	45	2	K
O	51-57	54	6	D,I,K
P	73-74	73.5	1	-
Q	80-83	81.5	3	E

km 28-33, and (M) in v(4) spans km 30-33; thus all three span and have centers within km 30-33 (table 2). The first two anomalies contain each others' centers, and so define a coincident pair. The anomaly in v(4) coincides with each of the other two anomalies. Thus, the three anomalies satisfy the extended rule that defines a triplet. However, suppose that the anomaly in v(1) were one km farther south, so that it spanned km 31-36. Its center would lie at d= 33.5. The anomaly in v(2) centers at km 30.5, so neither anomaly would contain the center of the other. The anomalies in v(1) and v(2) would no longer coincide. The triplet would then degenerate into two pairs, one linking v(2) and v(4), and another linking v(1) and v(4).

Identifying a quadruplet or larger coincidence of anomalies requires a more general rule, which also applies to triplets. For m (the number of coincident anomalies) greater than 2, an m-tuplet of coincident anomalies exists if and only if all possible coincident pairs of its component anomalies also exist. The number of possible coincident pairs is the number of ways that m anomalies can be chosen two at a time, or m(m−1)/2. A triplet requires three coincident pairs, a quadruplet requires six, and a quintuplet requires 10. This more general rule shows that the four anomalies at km 55 of figure 4 do define a quadruplet. Six coincident pairs create the quadruplet. However, if the anomaly in v(2) at d= 55 were one km farther south, then the anomalies in v(1) and v(2) would no longer coincide. The quadruplet would degenerate into two triplets, one linking v(1), v(3), and v(4), and another linking v(2), v(3), and v(4). Each of these triplets would contain three coincident pairs of anomalies, and so would not degenerate.

This extended rule for m-tuplets avoids a phenomenon that we call chaining. Chaining is illustrated by considering the hypothetical triplet that degenerated into two pairs. If a pair links v(2) and v(4), and if a second pair links v(1) and v(4), then a chain of two pairs links all three anomalies in these variables. However, the anomalies in v(1) and v(2) do not coincide. If the anomaly in v(4) were wide enough, the anomalies in v(1) and v(2) might even be several km apart. To accept such chained anomalies as a triplet would blur the idea of coincident anomalies beyond usefulness.

On cursory inspection of figure 4, v(1) and v(2) appear to be associated variables in the sense that most of their anomalies coincide. Variable 3 appears least associated with most other variables. However, the single anomaly of v(3) is part of the only quadruplet of coincident anomalies.

How much of the pattern described in the preceding paragraphs reflects some underlying structure that is worth trying to interpret, and how much should be dismissed as arising from chance? Two aspects of coincident anomalies are of interest: individual places on the traverse where unusually many anomalies coincide (anomalous sections of the traverse), and variables that are highly associated over the whole length of the traverse (associated variables).

Anomalous sections of the traverse differ in some way from adjacent sections. What this difference might be depends on what phenomenon the data were chosen to reveal. For the Wasatch fault zone, data were chosen to reveal segment boundaries, so anomalous sections of the fault zone could be segment boundaries. Highly associated variables might be more effective together than are single variables at identifying locations of the phenomenon that is under study. For example, if cross faults are thought to have localized mafic intrusions and volcanic rocks in an area, then coincident gravity and magnetic highs could identify hidden cross faults better than could either gravity or magnetic data alone.

ASSOCIATED VARIABLES AND JACCARD COEFFICIENTS

Associated variables are those that tend to have anomalies and lack anomalies together along the traverse, so that most of their anomalies coincide. Examples are v(1) and v(2) in figure 4. Association is observed, not inferred, and for now we make no genetic interpretations of any observed association. In contrast, unassociated variables are those that tend to have and lack anomalies without regard to each other. For example, in figure 4, v(3) appears to be relatively unassociated with v(1) and v(2).

The degree of association between pairs of variables that can take only values of 0 or 1 is measured by J, the Jaccard coefficient (Cheetham and Hazel, 1969). J measures the fraction of the anomalies that coincide. In a notation similar to that of Cheetham and Hazel (1969, p. 1131) J is defined as C/N(t). C is the number of coincident anomalies in each variable, that is, the number of sections of the traverse in which both variables have anomalies together. N(t) is the total number of anomalies present in both variables together, with each pair of coincident anomalies counted only once. That is, N(t) is the number of sections of the traverse where one or more anomalies are present. N(t) is calculated as N(i) + N(j)–C, where N(i) and N(j) are the numbers of anomalies in variables i and j. For example, variables 1 and 2 of figure 4 have five anomalies each, and four of them coincide. Therefore, C = 4, N(t) = 6, and J = 0.67. For most applications J = 0 for variables that have no coincident anomalies, and J = 1 for variables whose anomalies coincide completely (Cheetham and Hazel, 1969). Values of J that are near 0 characterize relatively less associated variables, and values near 1 characterize relatively more associated variables.

In most applications of the Jaccard coefficient, J cannot exceed 1. Then, by the definition of J, C cannot exceed the average of N(i) and N(j). In this application, if all anomalies in v(i) and v(j) are about equally wide, then probably no anomaly will be wide enough to coincide with more than one other anomaly. Then C is unlikely to exceed the average of N(i) and N(j), and J is unlikely to exceed 1. However, if some anomalies are much wider than others, then a wide anomaly might coincide with more than one narrow anomaly, so that C increases and J might exceed 1. For an example, consider v(3) and v(4) and only kilometers 40–70 of figure 4. The wide anomaly K in v(3) coincides with both narrow anomalies N and O in v(4). In this case N(3) = 1, N(4) = 2, and C = 2. C exceeds the average of N(3) and N(4), and J = 2/(1 + 2–2) = 2. However, the effect of the wide anomaly would be diluted if the entire traverse were considered. Then N(4) would increase to 6, so that C = 2 would no longer exceed the average of N(3) and N(4), and J would decrease to 0.40. Thus, J can exceed 1 if some anomalies are much wider than others, but such large values of J will be few if wide anomalies are few. In any case, large values of J will still identify highly associated variables.

We use J to examine the structure of the anomaly pattern of figure 4 (fig. 5, table 3). Variables 1 and 2 appear to be the most strongly associated (fig. 4), and the point (1,2) that represents this pair plots in the right part of figure 5 as it should. Variable 3 is comparatively unassociated with variables 1 and 2 (fig. 4). Accordingly, variable pairs involving v(3) should plot in the left part of figure 5 as they do. Plotting J in figure 5 appears to separate pairs of variables in ways that identify relatively more associated and relatively less associated variables.

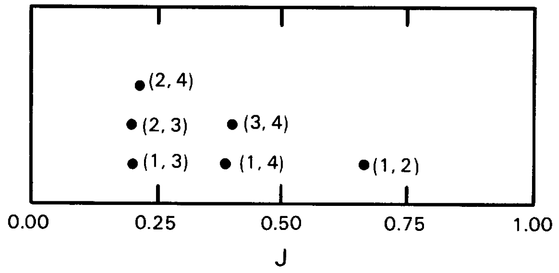

Figure 5. Illustration of use of Jaccard coefficient (J). Data are from figure 4. Calculations are summarized in table 3. Numerals to right of plotted points identify pairs of variables that produce each point. Vertical axis has no meaning and serves only to separate points for legibility.

Table 3. Calculation of Jaccard coefficients to produce figure 5 from figure 4

[C, number of coincident anomalies in a pair of variables; N(t), number of anomalies in both variables together, with coincident anomalies counted once; J, Jaccard coefficient, C/N(t)]

Variables paired	C	N(t)	J
1 and 2	4	6	0.67
1 and 3	1	5	.20
1 and 4	3	8	.38
2 and 3	1	5	.20
2 and 4	2	9	.22
3 and 4	2	5	.40

SIMULATIONS AND ANOMALOUS SECTIONS OF THE TRAVERSE

Simulations can help to identify sections of the traverse where anomalies in unusually many variables coincide. As explained earlier, the observed anomalies are identified and their widths determined by methods that depend on the data and goals of the investigation. However, once this is done, the number of anomalies in each data type and their widths are fixed. A simulation is performed by taking the observed number and widths of anomalies in each data type and placing the anomalies randomly along the traverse. The observed anomaly pattern is unlikely to have arisen by chance if some specified aspect of the pattern, such as a large number of coincident pairs, is found in the observed pattern but only in a small percentage of the simulated patterns. Any aspect of the patterns can be specified as the basis for comparing observed and simulated patterns. In this section we examine separately the number of pairs, the number of triplets, and the number of quadruplets of coincident anomalies. In a later section we shall compare observed to simulated patterns on the basis of all three numbers considered together. The basis for comparison must be specified before the observed and simulated patterns are examined to avoid distortion of results by prior inspection (Freedman and others, 1978, p. 494; Moore, 1979, p. 294-295; Wheeler, 1985).

The number and widths of anomalies in the observed pattern for each data type are held constant in the simulations. Only the locations of the observed anomalies are varied. The number of anomalies is held constant because each data type is measured reliably enough and its coverage is complete enough, that probably no anomalies have been missed. In terms of the variables v(i), this means that a value of 0 is as reliable as a value of 1. In fact, for any variable the locations of the 0's along the fault give exactly the same information as do the locations of the 1's, because either set of locations can be used to generate the other set. Anomaly widths are held constant because the edges of most anomalies are located with an uncertainty that is small with respect to the width of a typical anomaly, and small with respect to the variation in anomaly widths. Wheeler and Krystinik (in press) described how the data from the Wasatch fault zone meet these conditions of data quality.

As an example of this approach, 20 simulations were done using the data of figure 4. Randomly located anomalies were allowed to overlap the ends of the traverse, but could not overlap each other. The result was 20 anomaly patterns, each resembling figure 4 to some degree. Whether the observed anomaly pattern has arisen by chance can be determined by comparison to the 20 simulated patterns. Twenty simulations are too few to give conclusive results, but are enough to illustrate the procedure. In practice, many more simulations than 20 would be used. For example, for the Wasatch fault zone we used 300 simulations. The number of simulations that is needed can be chosen by an experienced statistician. Alternatively, the number of simulations used can be increased gradually, say in increments of 50 simulations, until results stabilize and cease changing much with each added increment.

The pattern that is observed in figure 4 has no more pairs or triplets of coincident anomalies than do many of the simulations (fig. 6). Despite the striking appearance of the triplet of figure 4, it would be unwise to spend much time trying to interpret it. The observed triplet is about as likely to reflect random processes as it is to betray the existence of any common cause of the three coincident anomalies. However, the observed pattern appears to be unusual in having a quadruplet. Only one of the 20 simulations has such a quadruplet. The observed quadruplet might merit investigation.

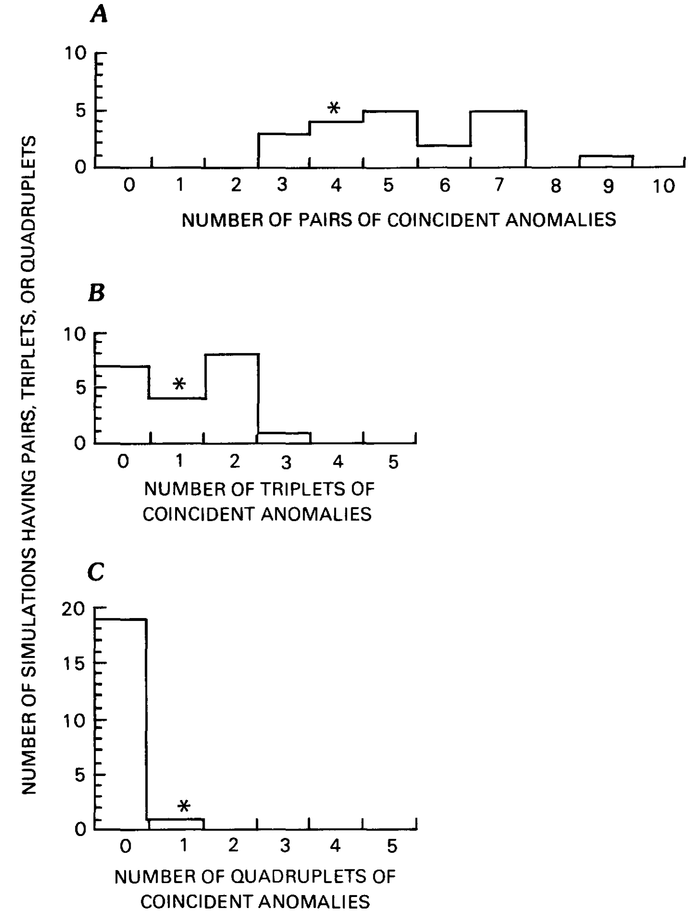

Figure 6. Histograms of numbers of pairs (*A*), triplets (*B*), and quadruplets (*C*) of coincident anomalies that occur in anomaly pattern of figure 4 (shown by asterisks) and in 20 randomized simulations of that pattern. Anomaly pattern of figure 4 has four pairs, one triplet, and one quadruplet. Thus, four simulated patterns contain same number of pairs (4) as does hypothetical observed pattern of figure 4. Four simulated patterns (not necessarily the same four) contain same number of triplets (1) as does figure 4. Only one simulation produced same number of quadruplets (1) as is observed.

RANDOMIZATION TESTS

Two-Dimensional Case

Triplets and quadruplets are of more interest than are pairs, because we seek places along the fault where several anomalies coincide. We do this because two randomly located anomalies are more likely to coincide than are three or four. A two-dimensional graph like that of figure 7 allows triplets and quadruplets to be examined together, instead of separately as with figure 6.

We examine the graph of figure 7 with a randomization test (Siegel, 1956; Conover, 1971; permutation test of Mosteller and Rourke, 1973) to determine whether the observed pattern of anomalies (fig. 4) contains more quadruplets and triplets of coincident anomalies than one would expect to occur by chance. Note that whole patterns are to be compared, not individual coincidences of anomalies. Interpretation of an individual triplet or quadruplet would require information that is not included in this randomization test.

To avoid distortion of the test result by prior inspection, the test and testable hypothesis must be formulated before examining the data. The population consists of the 20 randomized simulations of the observed anomaly pattern. The sample is the observed pattern. The goal of this randomization test is to tell whether the coincident anomalies in the sample are likely to have arisen from the same random processes that produced the population.

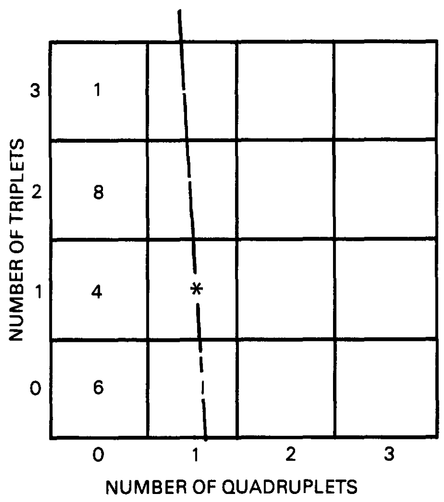

Figure 7. Numbers of coincident anomalies counted in anomaly pattern of figure 4 (shown by asterisk) and in 20 randomized simulations of that pattern. Pattern of figure 4 has one place along traverse where anomalies in three variables coincide and one place where anomalies in four variables coincide. Dashed line has a slope of $-t_s/q_s$, where t_s is total number of triplets in all 20 simulations together and q_s is total number of quadruplets. Dashed line is drawn to pass through the center of cell that contains asterisk and is used to determine whether asterisk plots significantly farther to upper right than do simulated patterns.

The alternative hypothesis, H(a), is one-sided, and states that the observed pattern was drawn from a population of patterns that typically have more triplets and quadruplets than occur in the simulated patterns. The null hypothesis, H(o), states that typical patterns in the population that yielded the observed pattern do not contain more triplets and quadruplets than do the simulated patterns.

The randomization test will produce a P-value (Moore, 1979; descriptive level of significance of Mosteller and Rourke, 1973; associated probability of Siegel, 1956, and Gibbons, 1976). The P-value is the probability of getting a simulated pattern at least as extreme as the observed pattern if H(o) is true. The extremeness of a pattern is defined as how far the pattern plots toward the upper right of figure 7. We will need a quantitative definition of "toward the upper right": in a direction perpendicular to the straight, dashed line in figure 7. To allow comparison to the observed pattern, the dashed line passes through the center of the cell that contains the asterisk. The dashed line has an orientation that reflects the approximate shape of the cloud of points from the simulated patterns.

Use of a straight line to approximate the shape of the point cloud in figure 7 requires justification, because this approximation influences the P-value of the randomization test. The choice of the slope of the dashed line also needs justification, because the slope influences the P-value. The purpose of the randomization test is to compare the observed anomaly pattern to the collection of simulated patterns, by comparing numbers of coincident anomalies (here, triplets and quadruplets). The number of triplets and quadruplets in the collection of simulated patterns must be summarized in some way in figure 7.

For the following reasons, the simplest though not necessarily optimal such summary is a straight line that slopes steeply down to the right or is vertical. First, random processes are more likely to cause anomalies in three variables to coincide than anomalies in four variables, so many simulations will have more triplets than quadruplets. Second, random processes are unlikely to produce a simulation with no triplets and no quadruplets, so comparatively few simulations will plot in the lower left corner of figure 7. Third, random processes are also unlikely to produce a simulation with many triplets and many quadruplets, so comparatively few simulations will plot in the central and upper right parts of figure 7. Fourth, the number of anomalies is fixed, so an increase in the number of triplets in a simulated pattern will usually mean a decrease in the number of quadruplets in that pattern. The result of these four reasons will usually be an elongated cloud of numerical entries in figure 7, sloping steeply down to the right and approaching or reaching the vertical in some cases. This elongated cloud of entries expresses the numbers of triplets and quadruplets in the various individual simulations. Because this cloud of entries is elongated, the simplest summary of its center is a straight line. The dashed line is drawn parallel to this center line to smooth out vagaries in the proportions of triplets and quadruplets in the individual simulations. The dashed line is offset to the right from the center line to pass through the center of the cell that contains the asterisk.

The optimal way to determine the slope of the dashed line is unknown, because the distribution that underlies the numbers of triplets and quadruplets in the simulations is unknown. However, a simple and straightforward expression for the slope is the ratio of the expected values of the numbers of triplets and quadruplets in the simulations, or E(t) and E(q), respectively. There are 20 simulations, which together have t_s triplets and q_s quadruplets. Therefore, the slope of the dashed line is -E(t)/E(q), which is estimated by $-(t_s/20)/(q_s/20) = -t_s/q_s = -23$.

The P-value is calculated as N(e)/N(p), where N(e) is the number of simulated patterns at least as extreme as the one observed and N(p) is the total number of simulated patterns. For the example of figure 7, N(p) = 20. No simulation plots on or to the upper right of the dashed line (the steeply sloping dashed line passes slightly to the right of the cell [1,0]). The P-value is 0/20 = 0. P-values from as few as 20 simulations can be unstable. If this P-value had arisen from many more than 20 simulations, the observed pattern would be concluded to have significantly more triplets and quadruplets than do the simulations. Then the triplet and quadruplet together could be interpreted as deserving further investigation.

Three-Dimensional Case

The example of figure 4 has four variables, so figure 7 ignores pairs and plots triplets against quadruplets. If there were five variables, we would need to replace figure 7 with a three-dimensional graph of triplets, quadruplets, and quintuplets against each other, again ignoring pairs. If there were six variables, we could ignore triplets as well as pairs and graph quadruplets, quintuplets, and sextuplets against each other. Alternatively, if the simulations for six variables contained few triplets, we might wish to consider them. If triplets are not abundant in the simulations, then sextuplets probably will be rare or absent. Then we could plot triplets, quadruplets, and quintuplets against each other, and could consider the few simulated patterns with sextuplets by adding them to the count for N(e). In any case, the replacement for figure 7 is unlikely to need more than three dimensions. The following discussion is in terms of triplets, quadruplets, and quintuplets; it could as easily be in terms of quadruplets, quintuplets, and sextuplets, or septuplets, octuplets, and nonuplets.

The problem is to recast figure 7 and its associated randomization test from two dimensions to three. Recall that the crux of the randomization test of figure 7 is to identify and count the simulated patterns that are at least as extreme as is the observed pattern, that is, the simulations that plot on or to the right of the dashed line that passes through the asterisk of figure 7. A more exact definition of extremeness could involve contouring the values of figure 7. Then, the simulated patterns that are at least as extreme as the observed pattern would be those that plot on or to the right of whichever contour passes through the center of the cell that contains the asterisk. For the example of figure 7, that contour would have a value of one, and two entries would plot on the contour, so N(e) = 2 and the P-value would be 2/20 or 0.1. Use of the dashed line gave a P-value of 0. Because the P-values are based on only 20 simulations the values might be unstable, but part of the difference between 0 and 0.1 reflects the difference between the two definitions of extremeness, that using the dashed line and that using contours.

At first glance it might seem that a three-dimensional version of figure 7, its dashed line, and its randomization test could be constructed straightforwardly from three two-dimensional figures and tests: triplets vs. quadruplets, triplets vs. quintuplets, and quadruplets vs. quintuplets. However, with much algebraic and geometric scribbling it can be shown that such an approach produces a result that is internally inconsistent. An alternative would be to contour in three dimensions. It is not clear to us how this could be done, either graphically or analytically. A remaining alternative is to construct directly a three-dimensional version of figure 7, its dashed line, and the associated randomization test (fig. 8, table 4).

In the two-dimensional case (fig. 7) the orientation of the dashed line was determined by the expected values of the numbers of triplets and quadruplets in the simulated patterns. The position of the dashed line was determined by requiring the line to pass through the point with the coordinates of the observed pattern. This point is shown in the figure by an asterisk. The randomization test involved identifying and counting all simulations that plotted on or outside the dashed line.

By analogy, the plane S′ and the point P of figure 8 correspond to the dashed line and the asterisk of figure 7, respectively. The orientation of S′ can be calculated from the values of t_s/n, q_s/n, and r_s/n, which estimate the expected values of t_i, q_i, and r_i. The position of the plane S′ can be calculated from the known coordinates of P. The randomization test will involve identifying and counting all simulations that plot on or outside S′ (on the side of S′ away from O).

Use of a planar significance surface S′ requires justification. The ith simulated pattern has coincident anomalies that determine values of t_i, q_i, and r_i. These three values define a point in three-dimensional space (fig. 8). There are n such points and they form a cloud of points. The significance of the observed pattern of anomalies depends on whether the point P lies farther from the origin O than does most of this cloud of points. Accordingly, we must construct a surface that passes through P and approximates the shape and orientation of this cloud of points. The plane S′ is such a surface, for the following reasons.

First, few if any simulated patterns will have so few coincident anomalies that they plot far inside S′, near O. Also, few if any simulations will have so many coincidences that they plot far outside S′. Most simulated patterns likely will have moderate numbers of triplets, quadruplets, and quintuplets, so that the cloud of points will tend to be thin in the direction perpendicular to S′.

Second, the number of anomalies that are available to coincide is fixed, so that any increase in one of t_i, q_i, or r_i requires a decrease in one or both of the others. This relationship will be strictly true if all anomalies are involved in triplets, quadruplets, and quintuplets. The

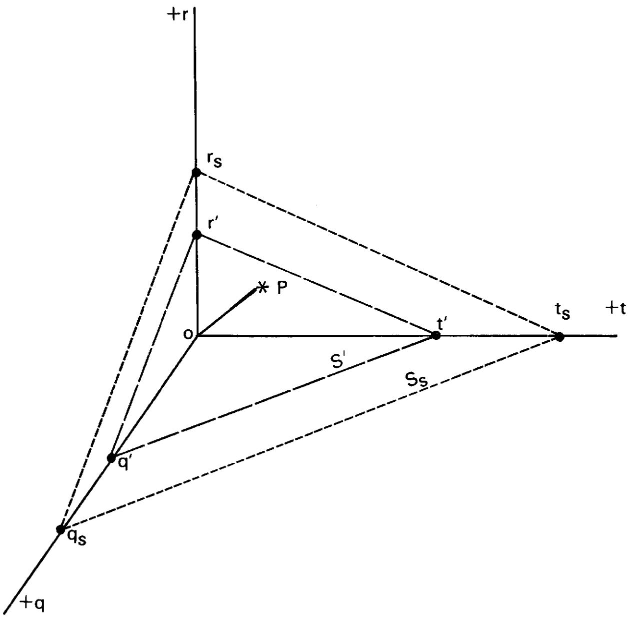

Figure 8. Sketch of oblique view of orthogonal coordinate system and points, lines, and planes used for three-dimensional case of anomalies of dissimilar widths. Table 4 defines symbols. Axes r and t lie in plane of page, axis q rises perpendicularly out of page. Taper of line from origin (O) to point P indicates that the line rises obliquely out of page toward reader. In general, line OP is not perpendicular to planes S′ or S_s. Point P is three-dimensional analogue of asterisk of figure 7, and is shown here as an asterisk to emphasize the analogy. For more than a few simulations $t_s > t'$, $q_s > q'$, and $r_s > r'$, so that S_s lies outside S′, that is, on opposite side of S′ from O. Randomly located anomalies are more likely to coincide to form a triplet than a quadruplet, and more likely a quadruplet than a quintuplet. Therefore, for more than a few simulations, the cloud of points from simulations will be distributed asymmetrically with respect to axes: $t_s > q_s > r_s$ and $t' > q' > r'$. However, single points need not share this asymmetry: it is not necessarily true that $t_o > q_o > r_o$, or that $t_i > q_i > r_i$ for any single simulation. Values of t_o, q_o, r_o, t_s, q_s, r_s, n, and (t_i, q_i, r_i; i = 1, ..., n) are known. Values of t', q', and r' are to be calculated.

Table 4. Symbols used for three-dimensional case, which has anomalies of dissimilar widths

Symbol	Definition
t	Number of triplets of coincident anomalies.
q	Number of quadruplets of coincident anomalies.
r	Number of quintuplets of coincident anomalies.
t_o, q_o, r_o	Values of t, q, and r, respectively, for observed pattern of anomalies.
t_s, q_s, r_s	Values of t, q, and r, respectively, summed over all simulated patterns of anomalies.
n	Number of simulated patterns.
t_i, q_i, r_i	Values of t, q, and r, respectively, in the ith simulated pattern, i = 1,...,n.
O	Origin of orthogonal coordinate system with axes t, q, r.
P	End point of a vector from O to (t_o, q_o, r_o).
S_s	A plane with t-intercept t_s, q-intercept q_s, and r-intercept r_s.
S'	A plane through P, parallel to S_s, and called the significance plane.
t',q',r'	t-, q-, and r-intercept, respectively, of S'.

relationship will also tend to be true if most anomalies are involved in such m-tuplets. The cloud will tend to spread out along the t, q, and r axes, so that the dimensions of the cloud in these directions will tend to exceed its thickness in the direction perpendicular to S′. The cloud will tend to be wide and thin.

Third, the wide, thin cloud of points will tend to undulate little, approximating a plane more than a highly curved surface. To see why, consider figure 8. For a constant number of quintuplets in a simulation, a quadruplet that loses one of its six component pairs will degenerate into three triplets. This conversion of quadruplets into triplets is roughly linear. Similarly, three triplets that are formed from five overlapping pairs can form one quadruplet if the missing sixth pair is added. This conversion of triplets into quadruplets is also roughly linear. Similar roughly linear relationships link triplets with quintuplets, and quadruplets with quintuplets. These considerations suggest that several simulations that plot in figure 8 with the same value of r (or of t, or of q) might lie approximately along a straight line that is perpendicular to the r axis (or the t axis, or the q axis) especially if most or all anomalies are involved in triplets, quadruplets, or quintuplets. The degree to which these lines perpendicular to the axes are straight is the degree to which the cloud of points approximates the plane S′.

Use of t_s, q_s, and r_s to calculate the orientation of S′ also requires justification. The expected value of the number of triplets in a typical simulated pattern is $E(t_i)$, which is estimated by t_s/n, and similarly for the numbers of quadruplets and quintuplets. Because multiplying the t-, q-, and r-intercepts of a plane by a constant, say 1/n, changes the position but not the orientation of the plane, a plane with intercepts $E(t_i)$, $E(q_i)$, and $E(r_i)$ is parallel to S_s and therefore to S′. Therefore, determining the orientation of S′ from the sums t_s, q_s, and r_s is equivalent to determining the orientation from the estimated expected values. This orientation will be slightly susceptible to the distorting effects of a few unusual simulated patterns, such as one with no coincident anomalies or one with a large number of quintuplets. However, if the number of simulations is large this distortion will be small and negligible.

The next step is to derive an equation for S′. S′ must be expressed in terms of the known values that are listed in the caption of figure 8. The intercept forms of the equations for S_s and S′ are, respectively,

$$\begin{aligned}(t/t_s)+(q/q_s)+(r/r_s)=1\\(t/t')+(q/q')+(r/r')=1\end{aligned} \qquad (1)$$

Because S_s and S′ are parallel, the coefficients of these two equations are proportional. If A is a constant,

$$(t'/t_s)=(q'/q_s)=(r'/r_s)=A \qquad (2)$$

Now P lies in S′, so by substituting the coordinates of P and equation (2) into equation (1), A is found to be

$$A=(t_o/t_s)+(q_o/q_s)+(r_o/r_s) \qquad (3)$$

Then, substituting equation (2) into equation (1) and multiplying through to clear fractions gives

$$tq_sr_s+qt_sr_s+rt_sq_s\text{-}At_sq_sr_s=0 \qquad (4)$$

Simulations will tend to have few quintuplets, so each simulation will plot in figure 8 as a point with an integral, usually small, value of r_i. (For example, each simulation for the data along the Wasatch fault zone has $r_i=0$, 1, or 2.) Then all points that constitute the population of the randomization test can be plotted in a few serial sections through figure 8 perpendicular to the r axis, each serial section lying at a different integral value of r (fig. 9). If the trace of S′ can be graphed in each serial section then the number of simulations that plot on or outside S′ can be counted easily by examining the serial sections. Setting r = k in equation (4) gives the equation of the linear trace of S′ in the plane r = k, k = 0, 1, . . ., as

$$tq_sr_s+qt_sr_s+(kt_sq_s\text{-}At_sq_sr_s)=0 \qquad (5)$$

The equations for the intercepts and slope of the trace are given in figure 10.

In terms of serial sections like those of figures 9–10, the P-value is N(e)/N(p). N(e) and N(p) are determined by counting the numbers of simulations that plot in various parts of the serial sections, analogously to the manner in which N(e) and N(p) were determined for the two-dimensional case by counting points in various parts of figure 7. N(e) is the number of simulations that plot

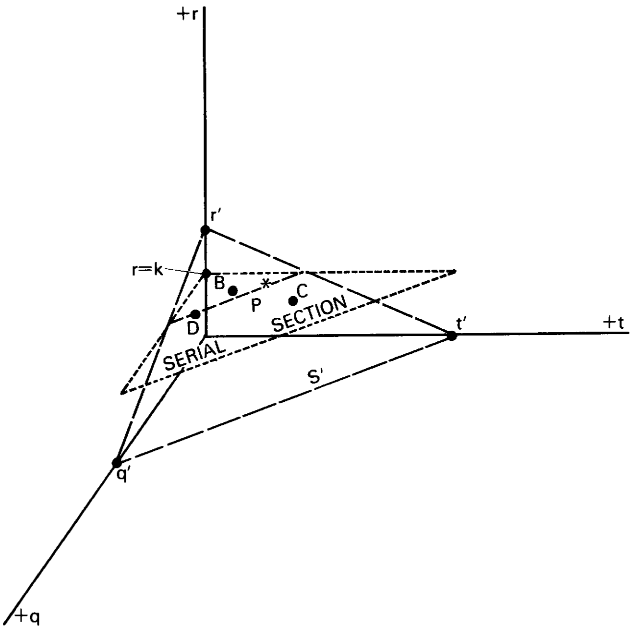

Figure 9. Part of figure 8, showing a serial section. Heavy lines and associated labels are from figure 8. Dotted lines outline a serial section that intercepts r axis perpendicularly at r = k, where k = 0, 1, 2, ... Only serial section through point P is shown. Other parallel sections lie at larger and smaller values of r. Serial section intersects plane S′ along short-dashed line which, in this section, passes through P analogously to dashed line of figure 7. Each simulated pattern that has k quintuplets plots in this serial section, either inside S′ (for example, point B), outside S′ (point C), or on dashed line (point D).

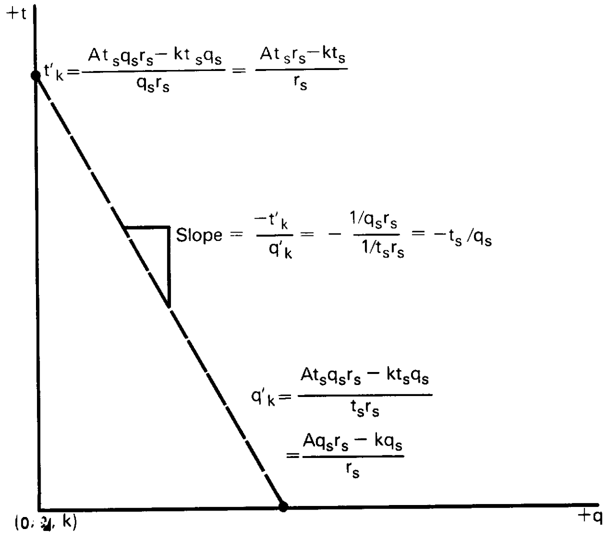
+t
$t'_k=\frac{At_sq_sr_s-kt_sq_s}{q_sr_s}=\frac{At_sr_s-kt_s}{r_s}$
Slope $=\frac{-t'_k}{q'_k}=-\frac{1/q_sr_s}{1/t_sr_s}=-t_s/q_s$
$q'_k=\frac{At_sq_sr_s-kt_sq_s}{t_sr_s}$
$=\frac{Aq_sr_s-kq_s}{r_s}$
(0, 0, k)
+q

Figure 10. Intercepts (t'_k and q'_k) and slope of trace (dashed line) of S′ in serial section r = k. Table 4 defines symbols; geometry is shown in figures 8–9. Value of A is given by equation (3). Origin of this serial section is not O of figure 8, but instead is point (0, 0, k) on r axis of figure 9.

on or outside the dashed lines, summed over all serial sections. In the example of figure 9, points C and D contribute to N(e). N(p) is the number of simulated patterns. In the example of figure 9, all of points B-D contribute to N(p).

SUMMARY

Associated variables can be identified by plotting J for the various observed pairs of variables (fig. 5). For the example of figure 4, only variables 1 and 2 are likely to be strongly associated along the traverse.

A pattern of coincident anomalies that is unlikely to have arisen by chance can be identified and evaluated visually by plotting counts of pairs and m-tuplets of coincident anomalies for observed and simulated anomaly patterns (figs. 6, 7). Numbers of triplets and quadruplets in the observed and simulated anomaly patterns are compared using a randomization test. If variables and simulated m-tuplets are too numerous to represent in two dimensions (fig. 7), a three-dimensional representation (figs. 8–10) should suffice. For the example studied here (fig. 4), results of the randomization test confirm suspicions derived from inspection of figure 7 and perhaps even from initial inspection of figure 4. The randomization test and inferences based on its result indicate that the observed pattern of anomalies contains more triplets and quadruplets than are likely to occur by chance. In practice, many more than 20 simulations would be needed to support statistical conclusions.

These conclusions are guides to aspects of figure 4 that are worth trying to interpret geologically. What interpretations are made depends on the natures of the variables and their anomalies. Further consideration of variables 1 and 2 together, and use of other geological or geophysical information, might suggest or confirm the existence of subtly expressed features at some or all of the places where variables 1 and 2 are anomalous together. The association of variables 1 and 2 along the length of the traverse indicates that these two variables together might be a more powerful investigative tool than is any of the four variables alone.

ACKNOWLEDGMENTS

W. F. Dummler wrote most of the simulation program and improved its design in numerous discussions. Comments by M.N. Machette and K.M. Shedlock improved the manuscript.

REFERENCES CITED

Cheetham, A.H., and Hazel, J.E., 1969, Binary (presence-absence) similarity coefficients: Journal of Paleontology, v. 43, p. 1130–1136.

Conover, W.J., 1971, Practical nonparametric statistics: New York, John Wiley and Sons, 462 p.

Freedman, Davis, Pisani, Robert, and Purves, Roger, 1978, Statistics: New York, W.W. Norton and Company, 506 p.

Gibbons, J.D., 1976, Nonparametric methods for quantitative analysis: Columbus, Ohio, American Sciences Press, 463 p.

Lewis, Peter, 1977, Maps and statistics: London, Methuen and Company, 318 p.

Machette, M.N., Personius, S.F., and Nelson, A.R., 1986, Late Quaternary segmentation and slip-rate history of the Wasatch fault zone, Utah [abs.]: Transactions of the American Geophysical Union, v. 67, no. 44, p. 1107.

Moore, D.S., 1979, Statistics—Concepts and controversies: San Francisco, W.H. Freeman and Company, 313 p.

Mosteller, Frederick, and Rourke, R.E.K., 1973, Sturdy statistics— Nonparametrics and order statistics: Reading, Massachusetts, Addison-Wesley, 395 p.

Ripley, B.D., 1981, Spatial statistics: New York, John Wiley and Sons, 252 p.

Schwartz, D.P., and Coppersmith, K.J., 1984, Fault behavior and characteristic earthquakes—Examples from the Wasatch and San Andreas fault zones: Journal of Geophysical Research, v. 89, p. 5681–5698.

Siegel, Sidney, 1956, Nonparametric statistics for the behavioral sciences: New York, McGraw-Hill, 312 p.

Wheeler, R.L., 1984, A plan for evaluating hypothesized segmentation of the Wasatch fault; *in* Hays, W.W., and Gori, P.L., eds., A workshop on "Evaluation of regional and urban earthquake hazards and risk in Utah", Salt Lake City, August 1984, Proceedings: U.S. Geological Survey Open-File Report 84-763, p. 576–605.

______1985, Evaluating point concentrations on a map—Earthquakes in the Colorado lineament: Geology, v. 13, p. 701–704.

Wheeler, R.L., and Krystinik, K.B., 1987a, Persistent and nonpersistent segment boundaries on the Wasatch fault zone, central Utah [abs.]: Seismological Research Letters, v. 58, no. 1, p. 31.

______1987b, Persistent and nonpersistent seismic segmentation of the Wasatch fault zone, Utah [abs.]: Geological Society of America Abstracts with Programs, v. 19, no. 5, p. 342.

______in press, Segmentation of the Wasatch fault zone, Utah—Geological and geophysical summaries, analyses, and interpretations: U.S. Geological Survey Bulletin, 180 ms. p.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1988—573-047/66,025 REGION NO. 8