Proper and Improper Use of Statistics in Geophysics*

By BLAIR KINSMAN, Chesapeake Bay Institute, The Johns Hopkins University, Baltimore, Maryland

(Manuscript received November 23, 1956, revised March 22 1957)

Abstract

A discussion of some of the peculiarities of the geophysical sciences together with their implications for statistical methods, particularly for the correlation coefficient, is given. The discussion is illustrated by an analysis of a recently published paper. In conclusion some advice on the use of correlations is offered.

Introduction

The notion that the business of a scientist is the telling of "likely stories" is at least as old as Plato's "Timaeus". The materials with which a scientist works are "appearances", that is sense impressions, and in the vivid old Greek phrase, he tries to "save the appearances". This is a fairly straight forward job. An excellent example is offered by Ptolemy, who takes astronomical observations back to the reign of Nabonasser and shows that he can "save" them, that is fit them into a coherent pattern, by telling either a geocentric story or a heliocentric story. With two equally satisfactory stories Ptolemy weighs their likelihood and, on the basis of the physical information available to him, develops the geocentric story as the more likely. The basis for this choice is hardly so simple and straight forward. Today the general consensus of opinion, except possibly among navigators, is that the heliocentric story is the "true" one.

The advent of the word "true" together with the word "real" in scientific discussion has done much to cloud the nature of scientific activity. The request for a "true" story instead of for a "likely" story tacitly postulates the existence of a "real" world underlying and giving rise to appearances and asks for information about that "real" world. The scientist *qua* scientist cannot answer such a question since the material on which he operates consists entirely of appearances. St. Thomas Aquinas would probably have said that no one could answer since "Nothing was ever in the mind that was not first begot upon the senses". The attempts to bridge the gap between the postulated "real" world and the world of appearances which we perceive has a long and uniformly unsatisfactory history covering the spectrum from Descartes' assertion that God would not fool us to Berkeley's retreat into solipsism. If we restrict ourselves to "appearances", "hypotheses", and "likelihood", metaphysical speculation about "truth" and "reality" can be left to the metaphysician with a considerable gain in clarity.

The point of departure is the appearances. These range from the casual impressions of any sentient being through the systematic observation of essentially uncontrollable phenomena characteristic of geophysics to the

* Contribution No. 31 from the Chesapeake Bay Institute.

precise measurement of the results of highly controlled experiments characteristic of the laboratory sciences. The habit of attentive observation, coupled with an overwhelming urge to fit the observations to a pattern, embryonic in most of the human race, is developed in the scientist to a high degree. All appearances, however, are not suitable for scientific activity. Aristotle said that the subject matter of science is that which happens always or for the most part. The unique event is a subject for history. POINCARÉ (1905) puts it this way: "Carlyle has written something after this fashion. 'Nothing but facts are of importance. John Lackland passed by here. Here is something that is admirable. Here is a reality for which I would give all the theories in the world.'.......... The physicist would most likely have said: 'John Lackland passed by here. It is all the same to me, for he will not pass this way again.'"

Having, then, a set of observations of a recurring phenomenon the next step is to construct an intelligible hypothesis into which the observations can be made to fit. One fertile method is the use of analogy. Some other set of phenomena and their pattern being known, if we can see a similarity, we can transfer the properties of the known system to the unknown. Since analogies are seldom isomorphisms, the correspondences being only partial, the dangers of argument by analogy are obvious. For example, the complex numbers are analogous to the real numbers in the sense that operations of addition, subtraction, multiplication, and division can be defined for each. We might then argue by analogy that, since division by zero is not permitted for the reals, division by zero is not permitted for the complex numbers. We thus reach a correct result. If we argue by analogy that, since the reals are ordered, the complex numbers must also be ordered our conclusion is false. Fertile as the argument by analogy is as a source of ideas, it is almost worthless in support of an hypothesis unless it is shown that the analogous systems are similar in every essential feature and that no dissimilar features can affect the properties that we wish to establish. Another method is to search the data for regularities. If the sample is small (and this is often the case in geophysics) this is quite easy to do. However, since even samples drawn from a random number table will exhibit regularities, results from such a procedure are suspect and once a regularity is found most scientists feel impelled to provide some rationalization for it, often based on an *ad hoc* selection of arguments. (The indiscriminate use of the correlation coefficient is particularly offensive in this connection and we will have something to say about it later.) Perhaps the most acceptable method of forming hypotheses is by rational argument from established elementary principles. It is worth noting that scientists in general seem to feel uneasy about any hypothesis until it has been presented in this form no matter how it was first conceived. An argument in Kepler's "Epitome" is a case in point. Kepler's hypothesis that the planetary distances are governed by the proportions of the five regular or Platonic solids seems a little wayward today, but the urge to order the welter of appearances is easily understandable. Kepler apparently felt the need to deduce his hypothesis from the first principles of geometry and astronomy since he devotes considerable space to the effort. His logic is impeccable except for one short section. There the line of the argument has been blurred, whether consciously or unconsciously cannot be known, so that the ostensible deduction he was at such pains to make is, in fact, not established. The expenditure of so much effort in such a cause by a man with first rate discoveries to his credit shows the importance attached to this method.

But the telling of tales is only half the job and the easiest half at that. We still face the problem of deciding how likely the story is or, if confronted with two different but adequate stories, which is the more likely of the two. Statistics has been increasingly concerned with understanding the structure of such decisions and with finding a clear and objective method of making them. The general problem is far from solved but many valuable results have already been obtained. The judgments of the likelihood of an hypothesis have had so many different bases that even a simple enumeration would be too long. They include decisions made on entirely extraneous grounds, e.g. the selection of a flat earth over a round earth on arguments derived from the "second coming". They include decisions made on what I should call aesthetic grounds, e.g. the selection of uniform circular motion as basic by the Greeks in contrast to uniform rectilinear motion by Newton. They include decisions in which maximum simplicity is equated with maximum likelihood. Occam's Razor is still a widely used scientific tool although the simplicity of nature is more an article of faith than a proven fact. The more we refine our studies of nature the more complex things become until it almost seems as though simplicity were an attribute of the infancy of a science. Fortunately, there are a few threads to guide us in the labyrinth. Occam's Razor has been mentioned. If you are willing to commit yourself to the proposition that the relations among phenomena are fundamentally simple, then you will always choose the simplest story that explains all the facts. However, it must always be borne in mind that one man's simplicity may be another man's utter confusion. A heliocentric hypothesis simplifies the astronomer's calculations but it makes those of the navigator intolerably complex. If one is willing to forego questions of "truth" and "reality" one can escape the dilemma by accepting both the geocentric and the heliocentric hypotheses (so long as they are not in logical conflict) and use whichever one is simpler *for the immediate purpose*. A most important criterion is that of compatibility with already existing structures. If the hypothesis under consideration would require extensive revision of major parts of a successful existing theory with all the labor that entails, clearly one would hesitate to accept it unless a general improvement throughout the whole theory could be anticipated. Another equally important basis for decision is the continued agreement of observations with the hypothesis. Here statistics enters since no set of experimental measures, if sufficiently refined, ever agrees exactly with an hypothesis or with other sets. The tincture of statistics that most of us retain from our formal training seldom goes beyond the memory of where to find a few computing formulae. This isn't enough. While a sprinkling of Pearson's x^2's, Student's t-tests, and Fisher F-tests do lend an air of objectivity to any paper, it must be remembered that statistical tests of significance derive from mathematical models which in turn are based on different views of the nature of phenomena. The calculation of parameters is routine and their use often obscures a lack of precise thought about the fundamentals of a problem.

In general, the job of the scientist is to invent a story which accounts for a set of observations and then to decide how likely the story is.

Geophysics

The geophysical sciences share the same theoretical base with the laboratory sciences. The Navier-Stokes equations hold equally well for a beaker of water and for the oceans. With this common base it is not surprising that the methods for treating geophysical data are often selected by analogy with those used on laboratory data. Unfortunately, the materials on which these methods are used in geophysics are sufficiently different from those of the laboratory sciences to require justification of the method, which is seldom explicitly given. In the first place, the equations which give a full description of any situation are usually too complex to be handled in their complete form. It is almost always necessary to simplify them by considering some terms as negligible in order to get an approximate solution. The laboratory scientist by controlling the conditions of an experiment can insure that terms considered negligible are so in fact and, as a result, he can expect good agreement between hypothesis and experiment even with relatively small samples. Further, he can repeat his experiment at will.

The geophysicist who, in the main can only observe, cannot repeat his observations in the sense of a repeated experiment. This is a grave difficulty since most statistical tests of significance are fundamentally rooted in the idea of repetition. While such repetition is conceptually possible in geophysics it is seldom realized and, in general, when such tests are used, their use should be supported by argument. Further, his inability to insure that factors considered negligible are so should logically force the geophysicist to use the equations for a process in their complete form or first to show that the neglected factors can be neglected. Neither of these courses is usually taken. There is nothing wrong with the making of simplified models so long as they are not offered as "reality". When the observations happen to agree with such models it is cause for gratification and suggests that the neglected terms were small. If the observations continue to agree with the model we can feel that we have a satisfactory story. In geophysics the agreement usually need not be very close before verification is claimed. In one case known to the author the originator of a widely accepted hypothesis will take a discrepancy of $\pm$ 20 % in the observations before he considers it worth any attention. This attitude stems from lack of control. In comparing data with a simplified model large dispersions are to be expected. This means that large samples are necessary if relations are to be established with any certainty. It is unfortunate when the investigator, having been forced to accept a simplified model, then feels impelled to insist that the natural phenomena are themselves simple.

Another difference arises from the answer that can be given to the question: "Do the data describe the phenomenon under consideration." In the laboratory sciences methods can usually be devised either to measure a property directly or to measure some closely linked property. With the method in hand, the experimenter can then accumulate enough data for a statistically reliable estimate of the property which interests him. The geophysicist here labors under two kinds of handicap. First is the matter of scale. Both the space and time scales are usually unwieldy. If one asks for the monthly average temperature of the Chesapeake Bay is it enough to dip a thermometer in once a day at some convenient place? One can hardly say without knowing a great deal about the structure of the Bay. If it were enough for, say, January 1949, would anything be known about January 1950? To answer such questions an inordinately expensive observation network would have to be established and maintained for many years. Salinity records taken daily at Solomon's Island, Maryland are a case in point. This set of data extending back to 1938 is the longest unbroken record of salinity taken anywhere on the Bay. Using monthly means and computing the power spectrum it was found that there was evidence of a yearly cycle, which was to be expected. However, the great bulk of the power in the signal occurred at periods greater than two years. It was calculated that to separate cycles having periods of three years and four years at the 5 % level, the record would have to extend over 285 years. Such a sample from a geophysical point of view is huge and it is rather sobering to see how little information it gives. Second, the geophysicist must frequently work with data which were taken for other purposes and which do not directly measure the properties that interest him. For example, an oceanographer interested in the factors influencing the size of fish populations might not have any measurements directly made for that purpose but instead measurements of salinity made at some point in the region and records of commercial fish catches. It isn't what he wants. It's what he's stuck with. If he persists, he would have to argue something like this. Fish population controls fish catch. Fish need plankton for food. Plankton need dissolved nutrients. Nutrients are brought to the surface layers by upwelling. Upwelling influences salinity. The salinity of the region can be determined from the salinities measured at a point which I know. Therefore, I will look over fish catches and salinities for possible correspondences and, if I can construct one, I will know the connection between fish population and environment. Laying aside the questions of whether the salinity measured at a point represents the salinity over a large area and of whether fish catch is an adequate measure of fish population, it seems unlikely that a definite clear cut relation between the ends of such a long and tenuous chain would emerge from a small sample.

Another hazard inherent in using existing data taken for other purposes arises from the temptation to fill gaps in it. If properties *A* and *B* are to be related and it is found that *A* was measured at some point for a number of years but that the measurement of *B* was neglected for a part of the time, then the urge to use measurements of *B* made somewhere else to fill the gap may be almost irresistible. This procedure enlarges the sample with an apparent increase in statistical reliability but it introduces tacitly the very difficult additional problem of showing that the measures introduced to fill the gap are the equivalent of what would have been secured had *B* been measured at the point. This is usually impossible. In using existing data for purposes for which they were not taken great care must be exercised to see that wishful thinking does not govern the make-up of the sample.

Geophysics needs large samples. It has small ones. In contrast with the laboratory sciences these small samples are often imprecise, having been painfully secured in the field over many years, sometimes two or three generations. To get another sample for testing involves the same long process. Thus, if the entire initial sample is used in the formulation of an hypothesis we are forced to leave its verification or rejection to our grandsons. It is clear that progress of a science which must either proceed on untested hypotheses or wait for generations to test them will be either insecure or very slow.

Verification is to be had only from data not used in formulating an hypothesis. One possible method of securing data for testing a geophysical hypothesis formulated by search, without delay, is to split the data on hand into two groups, one to be used in formulating the hypothesis and the other reserved for testing. This may be done in a number of ways. In some fields dealing with time series every other time unit is grouped to form the two sets, or the data may simply be split in the middle. Separation by means of some randomizing device could be used so that the bias, conscious or unconscious, of the investigator would not invalidate such statistical tests of significance as might be appropriate. The real difficulty here is "keeping the game honest". If the hypothesis is formed before the test data are taken no question of influence arises. With both sets of data in existence at the beginning of an investigation there is always the question of the extent to which the investigator is influenced in his selection of hypotheses by the test data. A glimpse of it, however fleeting, could bias him toward hypotheses likely to fit both sets. The difficulty could be met if the separation were made before the investigator saw the data and he inspected only one set until he was ready to test his hypothesis. Any alterations in the hypothesis after testing would, of course, be highly suspect. The advantage of this device is that verification can be carried out at once and an estimate of the value of the hypothesis made. The disadvantage is that the already small sample size is further reduced, but it may be worth accepting this reduction in exchange for immediate evaluation. It is well to remember that the information contained by any finite sample is limited. Manipulating it in this way cannot increase the amount of information contained. It can only sacrifice information of one kind to gain information of another.

Correlation

A fallacy that is still far too prevalent in geophysical papers is the inference from a high correlation to cause and effect. Such inferences, when they are made, are usually concealed under a mass of verbiage and intensive critical reading is usually necessary to reveal them for what they are. The existence of a cause and effect relationship is never a necessary inference from a correlation.

The exhibition of high correlations between unlikely things provides considerable amusement deriving from the absurd inferences that suggest themselves. One of the author's favorites is the high correlation between increases in school teachers' pay and increases in the production of alcoholic beverages from which one may conclude that the way to promote sobriety is to cut teachers' salaries. COHEN (1944) gives an interesting discussion of causality and correlation beginning on page 132 of his "Preface to Logic" in which he cites a correlation of 0.87 between the membership of the International Machinists Union and the death rate of the State of Hyderabad.

Correlations may be used legitimately for exploratory work. If an investigator suspects for any reason that properties A to D may have a causal connection with X and wishes to fix on the most likely ones for immediate study, the computation of the correlations may be used as a guide. Suppose that A is highly correlated with X and that none of the others are. Then a thorough study of the relation between A and X is an obvious first step. However, just from the nature of correlations the investigator may be misled. The size of a correlation in itself proves nothing. It is merely a signpost pointing the way and it may very well be pointing the wrong way. The term "high" correlation has been used without saying what is meant. In some fields a correlation greater than 0.60 seems high enough to cause excitement. How good is a correlation of 0.60? If we compute the coefficient of alienation we get 0.80 meaning that in comparison with a zero correlation, a correlation of 0.60 has reduced the standard error of estimate only 20 %. Eighty percent of the scatter still remains. The correlation must rise to 0.866 before the standard error of estimate is halved and even with a correlation of 0.99 the standard error of estimate is still 14 % of the standard deviation. As usually used in the literature, "high" seems to mean no more than "higher than we usually get with this kind of thing". It should be obvious that a correlation coefficient is not high for prediction purposes unless the major part of the variation is accounted for. As a rule of thumb one should not take a correlation seriously for prediction purposes unless it is 0.95 or higher, since it will then account for 70 % or more of the scatter *in the sample for which it was computed.*

The estimates of the true correlation in a universe computed from samples are strongly influenced by the size of the sample. If the true correlation is ϱ then the standard deviation of the sample correlations, r, is approximately $(1 - \varrho^2)/\sqrt{n-1}$. Thus, if n is small quite large departures of r from ϱ are fairly common. If $\varrho = 0$, then an exact test based on the t-distribution is $t = (r\sqrt{n-2})/\sqrt{1-r^2}$ with $n-2$ degrees of freedom. In dealing with small samples the greatest caution should be exercised in accepting a high sample r as evidence of a high population ϱ.

If correlations are to be computed, the data should be used without preliminary modification. Failure of the data to measure the properties under consideration will act to lower any correlation that may actually exist but, however uncertain are the original data, no extra information can be forced into them by sitting at a desk and manipulating them. Occasionally one encounters a case in which the data have been first rounded off. The process of rounding off is sometimes useful as a means of expressing the degree of doubt about the original data, but calculating correlations with rounded figures may lead to biased and imprecise results. Less often one finds the data grouped and correlations computed between arbitrary numbers designating the groups. The worst of all practices, however, is the artifice of adjusting the data to improve the correlation. If the values are adjusted, the correlation can be improved to almost any extent but it then has little bearing on the phenomena the data purport to measure.

Another practice sometimes used with a pair of vaguely sinusoidal records such as frequently derive from meteorological phenomena is to slide them past each other until a fair correlation is obtained. Further lags are then investigated until another fair correlation is obtained and this is then used in connection with the first one to improve the agreement. Since this can always be done with truncated functions of this sort, the improvement is probably factitious.

If a correlation were computed and tested and allowed to stand or fall by itself the situation would not be too confused. However, one often finds that whole series of correlations have been computed. In general when a number of possible correlations are available it is not very fruitful to fix attention on one impressive one and ignore the others. This process of selection is not always deliberate. The chapter on ESP and PK in GARDINER (1952) illustrates how it may operate without any intention to deceive on the part of the investigator.

The correlation coefficient computed for a sample, then, is evidence for a relation obtaining within the sample for which it was computed. It is an estimate of an unknown conceptual true correlation existing in the universe from which the sample was drawn but it is an estimate that may easily be far wide of the mark. It is certainly no convincing evidence of any cause and effect relation. From the discussion, it should be clear that while correlations can be useful as supporting evidence, they can never play the role of primary arguments. In fact, the entire field of geophysics would be spared a great deal if every author would withhold from publication any paper devoted mainly to presenting arguments constructed to account for a high correlation.

Icebergs

The foregoing general discussion had its genesis over several years during which time critical reading of many papers in geophysics and allied fields produced the uneasy feeling that quite a few of them contained less than met the eye. It was precipitated specifically by a paper by LOUIS A. POST (1956) called "The Role of the Gulf Stream in the Prediction of Iceberg Distribution in the North Atlantic".

After stating the fact that the number of icebergs south of 48° N per year varies widely, Post rejects the commonly held opinion that the causes are many and related in a complex fashion in favor of the notion of a single primary cause. He states his objective as follows: "It is the purpose of this paper to provide evidence that fluctuation in the Gulf Stream System is such a single cause and to derive a prediction technique for icebergs south of latitude 48° N from causative variates in this system."

We consider first the argument for a primary cause. When this is disposed of we will still have to look into the effectiveness of the prediction system whether the cause exists or not, since it is conceivable that one variable may be a good predictor of another even though they are not causally related.

Post's argument may be summarized as follows:

I. The Gulf Stream System, together with the Irminger, West Greenland, and Labrador Currents, may be taken as analogous to a closed circulatory system, information secured at one point being applicable to other points at other times.

II. The mean monthly temperature anomaly of the Florida Current at Key West[1] is inversely proportional to the "strength" of the current at Key West.

III. The "strength" of the Gulf Stream off the Grand Banks varies inversely as the temperature anomaly at Key West after a suitable delay.

IV. As a result of I through III, the number of icebergs south of 48° N varies directly with the anomaly.

V. After leaving the Grand Banks some of the water from the Gulf Stream is carried into the iceberg calving area in Baffin Bay by the Irminger and West Greenland Currents. Here a strong influx of warm water increases the number of icebergs calved so that a strong Gulf Stream current will increase the number of icebergs available for transport by the Labrador Current. It also serves to strengthen the Labrador Current so that after a suitable lag a larger number of bergs appears in the Grand Banks area. Since a strong Gulf Stream is associated with a low anomaly, the variation in icebergs with anomalies is inverse, allowing for a sufficient time lag to permit the current to reach Baffin Bay and return.

The analogy with a closed circulatory system (I) provides an appealingly simple picture. However, the absence of restraining walls in the Gulf Stream System makes it unclear that effects are transmitted through it in a manner analogous to their transmission through a closed system of tubes.

In the absence of velocity or mass transport measurements at Key West during the period under consideration, the argument in support of assertion II, that Key West temperature is inversely proportional to Key West current "strength", must be indirect. The data used as a test are the counts of icebergs found south of 48° N made by the International Iceberg Patrol[2] and the mean monthly temperatures at Key West (UNITED STATES DEPARTMENT OF COMMERCE, Special Publication No. 278, Revised (1951)). A correlation of 0.65 is found between the March anomalies and the iceberg count for 26 pairs of measures. This is evidence of a sort for a direct variation of the number of *icebergs* with the March anomaly. It is worth noting that the iceberg counts on which the correlation is based are highly subjective and have been made by many different people. In addition there is reason to suppose that the very low counts recorded for 1940 and 1941 may have been the result of lack of observers rather than lack of icebergs. If a connection does exist the error in the iceberg count would act to decrease the estimate of the correlation coefficient. If one accepts Post's contention that the number

[1] Hereafter, the word anomalies will denote mean monthly temperature anomalies of the Florida Current at Key West unless otherwise specified.

[2] The primary sources for iceberg counts are The Bulletins of the U.S. Treasury Department, Coast Guard, "International Ice Observation and Ice Patrol Service in the North Atlantic Ocean". Bulletin No. 21, the Season of 1931, tabulates counts from 1900—31. Bulletins Nos. 22—34 cover the years from 1932—48. The counts for years subsequent to 1948 were furnished to us by the U.S. Coast Guard, Washington, D.C. at our request. Whether these are the counts used by POST (1956) is not known.

of icebergs varies inversely with the "strength" of the Gulf Stream, it follows with the same force that the "strength" of the Gulf Stream varies inversely as the March anomaly. The step from the "strength" of the current off the Grand Banks to the "strength" of the current in the Straits of Florida is not very firm. Two arguments are offered. One, based on ISELIN (1938, 1940) holds that positive anomalies are associated with a stronger Florida Current which brings about a contraction of the North Atlantic Eddy. This in turn is accompanied by weaker peripheral currents in the region of the Grand Banks which permit the Labrador Current to carry its load of icebergs farther to the south thus increasing the number counted. The other, arguing by analogy with a closed system, holds that weak currents are transmitted directly through the system so that weak currents in the Gulf Stream imply weak currents in the Straits of Florida. The second argument is preferred. The assertion is thus seen to hinge on a correlation coffiecient, an assumed relation between Gulf Stream "strength" and numbers of icebergs, and an argument by analogy. The reasons for the selection of the March anomalies will be discussed later, but it should be noted here that three other months show higher correlations, July and December of the previous year, 0.72 and 0.70 respectively, and January, 0.76. Too much weight should not be given to the differences between these correlations as the sample size is not large enough to rule out substantial chance fluctuations. In the absence of data the assumed relation between Gulf Stream "strength" and numbers of icebergs is unsupported. The argument by analogy, while a useful guide, cannot serve as a secure foundation for a theory of causation. The contribution of the Antilles Current to the Gulf Stream has not been considered. A substantial part of the mass transport of the Gulf Stream has been contributed north of the Florida Current. Fluctuations in this additional flow might not agree with those of the Florida Current and the fluctuations in the Gulf Stream might be markedly different from those of the Florida Current. There is little direct evidence. Two studies by WERTHEIM (1953, 1954) provide some indirect evidence that the action of the Antilles Current may frequently be that of a damper on the Florida Current.

Assertion III, that Gulf Stream "strength" off the Banks varies inversely as Key West temperature anomaly, again goes back to the correlation. It is assumed that a weak Gulf Stream current off the Grand Banks permits the Labrador Current to carry its load of icebergs farther to the south while a strong Gulf Stream sweeps the icebergs northward out of the area. If this is granted, then the correlation of 0.65 between March anomalies and iceberg numbers implies an inverse variation of the Gulf Stream "strength" with temperature anomaly at Kay West.

The conclusion, IV, that the number of icebergs south of 48° N varies directly with the March temperature anomaly at Key West, is not a result of the previous steps. It is a direct inference from the correlation of 0.65. From the fact that the Gulf Stream system is the major source of warm water in the region of the Grand Banks it is reasonable to expect some connection between temperature fluctuation at Key West and iceberg numbers. That the variation is direct is suggested by the fact that the correlations for each of the twelve months is positive.

Table 1. Correlation of Iceberg Numbers and Mean Monthly Temperature Anomalies at Key West

Month	June	July	Aug.	Sept.	Oct.	Nov.
Correlation ..	0.31	0.72	0.08	0.07	0.17	0.10
Month	Dec.	Jan.	Feb.	Mar.	Apr.	May
Correlation ..	0.70	0.76	0.15	0.65	0.30	0.40

That the temperature anomaly at Key West is a primary cause is unproven. Many phenomena show fluctuations similar to the anomalies at Key West and could be made, by the type of selection used by Post, to show correlations with iceberg numbers as high or higher. In casting about for something so remote from iceberg numbers that there would be no temptation to fabricate an *ad hoc* argument for causation the author hit upon the notion of correlating the typeset commas per page in Post's paper as it was printed in Tellus. After a few minutes of exploring it was found that by pairing pages 111—102 with iceberg counts from 1942—1951 a correlation of − 0.81 resulted. This is considerably better than 0.65 but no one would argue that commas cause icebergs. GARDINER (1952) gives an interesting popular discussion of the ease with which such striking correspondences can be manufactured in his chapter on the Great Pyramid.

The selection of the March anomalies needs some comment. It is stated that"... three months is the maximum time lag for current between the Straits of Florida and the Grand Banks". Considering the paucity of information and the turbulent nature of the flow it is difficult to attach a precise meaning to this statement. It may reasonably be taken to mean that some considerable but unknown fraction of the water particles passing Key West will be found off the Grand Banks in not less than three months. From the records of the Ice Patrol, on the average 87 % of the bergs occur in the period from March to June. If the time from Key West to the Grand Banks is correctly estimated, then anomalies for the months from January through May might be important, with those for March likely to arrive at the height of the iceberg season. All questions of physical significance aside, the correlations for two of these months fail to attain statistical significance at the 0.1 level, one of them is just above the 0.05 level, and the other two are above the 0.001 level. In general, when a number of possible correlations are available it is not very fruitful to fix attention on one impressive one and ignore the others. The "high" correlation for March looks very much like an accident.

An even greater difficulty in accepting the March correlations as meaningful arises from the fact that temperature measurements at Key West are lacking from 1941 through 1945 and the gap has been filled with temperature anomalies measured at Daytona Beach. Since the sample contains only 26 pairs of measures, the interpolated 5 represent 19 % of the data and possible error from this source must be considered. There are 118 months in which temperatures were measured at both places. The standard deviation of the differences between the temperature anomalies at Key West and at Daytona for these 118 pairs is $\pm$ 2.21 while the standard deviation of the 21 measures actually available from Key West is $\pm$ 2.29. From this it can be inferred that the dispersion of the errors introduced with the Daytona data is probably of about the same size as the dispersion of the Key West measures themselves.

Actually a correlation of 0.65 is not very high. The coefficient of alienation is 0.76 so that the standard error of estimate is reduced only 24 %. For example, if the mean value of the berg counts is used to predict, the standard deviation is $\pm$ 360 bergs. If the anomalies having a correlation of 0.65 are used, the standard error of estimate is $\pm$ 275 bergs, a reduction of about 24 %, so that the quality of the estimate is still poor.

The uncertainty of the iceberg data has been mentioned. "Surface water temperatures at tide stations, Atlantic Coast, North and South America" says the temperature data "are based upon thermometer readings made in a sample of water drawn by bucket from a foot or two below the surface. The observations are usually made once each weekday at whatever time the observer tends the tide gage..." That the anomalies secured from these data are representative of the fluctuations in the temperature of the entire Florida Current may be doubted so that, even if the whole picture presented by Post were correct, it is questionable whether the numbers he uses do in fact describe the phenomenon about which he wishes to talk. This objection is quite aside from the question of how well the numbers may serve to predict.

From current measurements the time lag for the delayed effect, V, is set at three years and a correlation of -0.76 is found. If the argument was developed before an inspection of the data, this correlation looks impressive. However, it seems that the time lag of three years inferred from current data is much more doubtful than the estimate of one to three months for the transit from Key West to the Grand Banks. The series of correlations for lags from one to eight years which were presented permit the suspicion that the theory was influenced by the correlations.

We thus arrive at the following conclusions:

1) That the appealingly simple picture of the mechanism connecting iceberg numbers with temperature anomalies at Key West is correct seems quite unlikely.

2) That fluctuation in temperature at Key West is a single cause capable alone of causing the greater part of the fluctuation in iceberg

count is unproven and seems highly unlikely.

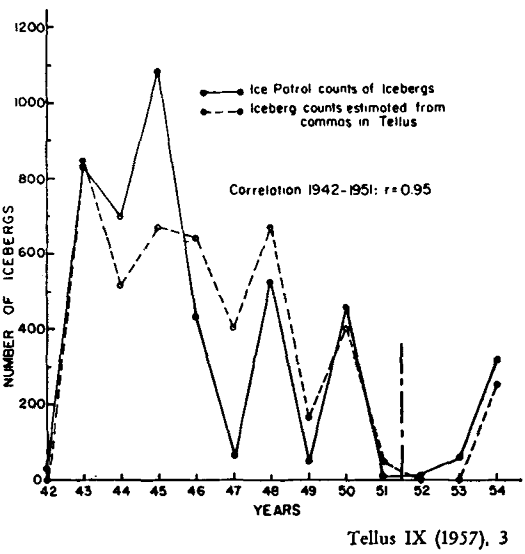
The proposed prediction method fares no better under close scrutiny than the hypothesis. The base of the method is a plot of iceberg counts on a grid of temperature anomalies, one axis being the anomaly of the same year as the iceberg count and the other axis the anomaly of three years before. (POST, 1956, Fig. 6.) There are 23 available iceberg counts. In plotting, the counts have been rounded off as shown in Table II.

Actual International Iceberg Patrol Counts	Actual Counts Accordng To Post's Eigure 6
2	0
2	0
6	0
13	0
22	0
47	0
63	0
30	30
216	225
?	250
430	430
460	475
470	475
475	475
514	500
523	525
575	550
664	670
700	700
840	840
850	850
872	875
1,087	1,075
1,351	1,350

The system on which rounding off is based is not apparent, particularly for the low values. One count too many has been plotted and given a value of 250. The plot made, it is contoured into four zones of approximately equal width, three of which represent a range of 200 bergs while the fourth has a range of 300 bergs. There is no indication of the band to which the dividing numbers 0, 200, 500, and 700 are to be assigned. Inspection shows that the contouring here is extremely unorthodox. Ten out of the 24 counts shown are included in the wrong zones and many of the misplaced counts are not "near misses". This incredible plot seems to represent the entire substance of the prediction method since all subsequent graphs and tables are derived from it. Post's Figure 7 purports to be constructed by reading values from Figure 6 but there are serious disagreements in the values shown in the two figures as was substantiated by the readings made independently by five qualified men.

It is asserted that "a coefficient of +.94 was obtained between computed and observed numbers grouped as in Figure 7 which allows a leeway of ± 100 bergs". The "Iceberg Prediction Table", Post's Table I, says, "For iceberg = 100 correlation coefficient is +94 (see Figure 7)". The process intended is not sufficiently clear to permit a definite check. If one accepts the values shown in Post's Figure 7 and correlates them, the result is 0.83. If the zones as taken in Post's Figure 6 are numbered from 1 to 5 lumping the "No iceberg" region with the 0 to 200 berg zone as 1, the right of assigning borderline counts in the most advantageous way is reserved, and the actual counts are rounded by anything up to ± 100 in such a way as to force the greatest degree of agreement, then the zone number correlation can be forced up to 0.92. Some such purely artificial method which has little relation to the original data seems to have been used.

Verification for the predicting method is claimed for the past three iceberg seasons and *Newsweek Magazine*, July 13, 1953 and *Science Newsletter*, May 1, 1954 and April 30, 1955



are cited, but no criterion by which agreement was judged is offered. As is well known, short runs can be highly misleading. The reader is referred to Figure 1 which shows icebergs as correlated with commas in *Tellus* 8, pages 102—111. The years 1952—1954 were not included in the original analysis and can therefore be considered as "prediction". As can be seen the agreement is much the same as the period 1942—1951. If the comma "prediction" is applied to the iceberg counts for 1937—1941, however, it fails completely, producing an average miss for the five tries of 544 icebergs. Figure 1 should be compared with Post's Figure 7. There is little reason to suppose that Post's predictor will hold up over any long period based as it is on a highly imaginative and unorthodox piece of contouring which fails to "predict" very well even the counts on which it is based.

Conclusion

In conclusion a few cautionary suggestions may be offered on the use of correlations. Correlations may be used freely in the search for possible connections among phenomena and are well defined and meaningful for the samples for which they are calculated. However, inferences from a sample to the population from which it was drawn should be made cautiously with due consideration for the influence of sample size and for the devaluation inherent in selecting one high correlation from a series. Data should never be manipulated to produce a "high" correlation and cause and effect should not be inferred from correlation alone, although correlation may be used as support for other arguments. It would be well if authors would withhold from publication papers whose arguments are constructed to account for a high correlation.

REFERENCES

COHEN, M. R., 1944: *A Preface to Logic.* New York, Henry Holt and Co., pp. 132—134.

GARDINER, M., 1952: *In the Name of Science.* New York, G. P. Putnam's Sons, pp. 173—185, 299—311.

ISELIN, COLUMBUS O'D., 1938: "A Promising Theory Concerning the Causes and Results of Long Period Variation in the Strength of the Gulf Stream System." *Trans. Amer. Geophys. Union*, **19**, pp. 234—244.

— 1940: "Preliminary Report on the Long period Variations in the Transport of the Gulf Stream System." *Papers in Physical Oceanography and Meteorology*, **8**, No. 1, pp 1—40.

POINCARÉ, H., 1905: *Science and Hypothesis.* New York, Dover Publications, Inc., p. 141.

POST, LOUIS A., 1956: "The Role of the Gulf Stream in the Prediction of Iceberg Distribution in the North Atlantic." *Tellus* **8**, pp. 102—111.

UNITED STATES DEPARTMENT OF COMMERCE, Coast and Geodetic Survey, "Surface Water Temperatures at Tide Stations, Atlantic Coast, North and South America." Special Publication No. 278, Revised (1951).

WERTHEIM, G. K., 1953: "Studies of the Electric Potential Between Key West, Florida and Havana, Cuba." *Woods Hole Oceanographic Institution*, Reference No. 53—95 (Unpublished manuscript).

— 1954: "Studies of the Electric Potential Between Key West, Florida and Havana, Cuba, No. II." *Woods Hole Oceanographic Institution*, Reference No. 54—68 (Unpublished manuscript).

Fig. 1: Comparison of iceberg numbers from 1942 to 1954 as counted by the International Ice Patrol and as estimated from commas in *Tellus*. (1942—1951 sample; 1952—1954 "prediction".)