A NOTE ON THE USE OF PATH ANALYSIS IN ANALYZING AND INTERPRETING OBSERVATIONAL DATA, WITH REFERENCE TO THE ANALYSIS OF GOOSE KILL AROUND LAKE MATTAMUSKEET, NORTH CAROLINA, 1960-1961 *

By W. Scott Overton, *N. C. State College, Raleigh*

and

Otto Florschutz, *N. C. Wildlife Resources Commission*

INTRODUCTION

Most wildlife and fisheries research involves the study of a system, either ecological or socio-ecological, and the data are observational in nature. By this is meant that no experimental control is exercised over the sampling units, but rather that natural phenomena are observed and recorded as they occur. (It is hoped that the sampling methodology is such that valid inferences from the sample to the population are possible.) Now to say that no experimental control is exercised is not in itself a condemning comment, as much can be learned from observing natural phenomena as they occur. The problem is one of properly interpreting these observational data, and in forming hypotheses that are truly consistent with the data.

In the present paper, we discuss and illustrate the distinction between observational data and the experimental data which characterizes many subject matter fields, and the differences in analytic and interpretive processes necessitated by this distinction. The method of Path Analysis (Turner and Stevens, 1959; Tukey, 1954; Wright, 1934, 1960; and Mallios, 1961) is presented as an aid in the analysis and interpretation, and also as a helpful tool in defining and describing the nature of the system under study.

DEFINITIONS

Communication is always a problem in presenting a topic such as this, so we will begin this paper with a list of definitions. These should not be construed to be absolute, in the sense of establishing these definitions for these terms throughout the field, but rather they should be considered an explanation of what *we* mean by these terms in *this* paper. Also, these fail to be absolute in another sense, as they are rather sketchy, and will be supplemented by discussions throughout the paper.

Path. The term path is used to indicate a functional relationship between two variables. The direction of the path indicates the direction of the cause.

Variable. A variable is any measurement, observation or characteristic associated with the observation units, sampling units, experimental units, or analytic units under study.

Observation unit. An observation unit is the smallest unit for which a distinct set of variables is observed. In the present case, this is a goose field for one day, as no records are available for the individual hunters using the fields, rather only the totals for all hunters per field.

Sampling unit. The sampling unit is the set of observation units that are defined to be a unit (selected together as a unit) by the sampling scheme. As all days and virtually all fields are included in the present study, we will have no need for this term in the present analysis.

Experimental unit. The group of material (or observation units) to which a "treatment" or level of "treatment" is assigned at random or in accordance with some experimental design.

Analytic unit. The largest collection of observation or experimental or sampling units for which common variables or treatments are defined by the analysis and by the sampling scheme or experimental design. In the present case,

* A joint contribution of the Southeastern Cooperative Fish and Game Statistics Project and the North Carolina Wildlife Resources Commission, Federal Aid Project, W-6-R.

this is the collection of all fields for which data are available on a given day.

Experimental data. Data such that the treatments or levels of treatment have been assigned at random to the experimental units.

Observational data. Data such that the variables of interest occur in a natural manner, without experimental restraint, over the observational units.

Regression coefficient. The linear regression coefficient of one variable on another. This is the observed change of the first variable for a unit change of the second, with no necessary inference of cause.

Path coefficient. The causal linear regression coefficient of one variable on another. If this is designated by *"a"*, it is indicated that the change of one unit in the second variable results in the change of *a* units in the first variable.

OBSERVATIONS ON A NATURAL SYSTEM AND PATH ANALYSIS

In analysis of data pertaining to a system, we are usually concerned with, (1) discovery of functional relationships within the system and, (2) description of the functional relationships. In turn, description involves both model building, or determination of the form of the relationship, and estimation of the parameters of the model. As the hypothesized model is involved in the definition of tests needed in the discovery phase, it is seen that the two phases are not entirely separate.

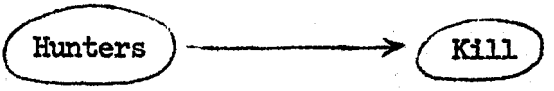
Analysis begins with the detection of association between elements in the system. This may be very elementary, as when a field man notices that a population has responded differently in a particular year or area, and then begins a search for some possible causative force which is also different in that year or area. He is utilizing what we may call the principle of correlated variation. This is practically instinctive to research people, and is the foundation for most elementary statistical techniques such as analysis of variance, analysis of covariance, correlation and regression. In these techniques one measures the variation in the variable of interest, Y, and the amount of this which is associated with the variation in a co-variable, X.

It is then a simple extension of these ideas to examine the variability left in Y after accounting for that associated with X^1, to see if any of this remaining variation can be associated with a second variable, X^2. For example, if one were studying the age-weight characteristics of a deer herd, it would be natural to first account for sex differences. Similarly, in studying the sex-weight differences in a deer herd, one should certainly take into account the relationship of weight on age. These examples can be handled readily by fairly elementary statistical techniques, and by now some of you are probably wondering what we are building up to.

Suppose you found after drawing your conclusions with regards to the weight difference between bucks and does that you actually had data from two areas, one of which was overpopulated and open to doe hunting, and the other with a population below carrying capacity and open only to bucks. It would then be necessary to modify the analysis and draw new conclusions based on this new information. The simple association relationships within these data would be very misleading, and good analytic results would follow only the analysis of the complete "system." This pretty well states the case of all observational analysis: the *complete* system must be considered, or there is risk of invalid conclusions. In an experiment, in contrast, the act of randomization allows one to confine attention to parts of the system that are of particular interest.

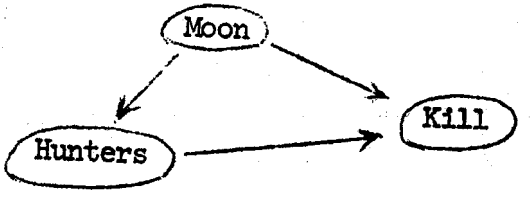
How, then, does one study a complete system? The method of path analysis is very helpful in studying biological systems and we will illustrate its use in the analysis of goose kill data from the fields around Lake Mattamuskeet, North Carolina, during the 1960-1961 hunting season. The Wildlife Resources Commission was able to get complete reports from almost all of the goose fields in this area for each day of the season. The data are total reported hunters and total reported geese killed, by day. Other variables considered are moon, day of season, goose population, and several weather variables. We will discuss these in some detail as we develop the path representation of the system.

Consider the following diagram,



(1)

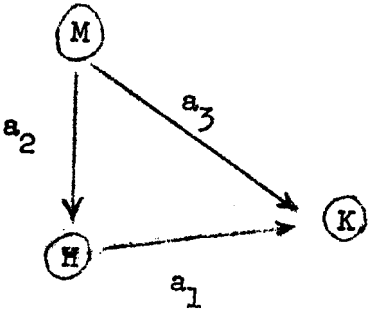
which says that hunters cause goose kill. There cannot be much argument about that, but if one wishes to study the rate at which goose kill changes as a result of a change in the number of hunters, then difficulties arise. For example, it is well known that geese will feed on a moonlit night, and generally accepted that this reduces their activity during the day and hence availability to the gun. In fact, this is so well accepted that hunters plan their trips accordingly. That is, we have the following path diagram when considering kill, hunters and moon.



(2)

Diagram (2) says that Kill is influenced by the Moon and the Hunters, and that Hunters are in turn influenced by the Moon. Therefore, the line of force from Moon to Kill is split into a direct force and a force through Hunters. It is clear that this diagram also is a gross oversimplification of the true state of affairs, but before proceeding with the evolution of a more complete path representation of the system, it will perhaps be worthwhile to discuss (2) as though it were itself complete.

We shall make a further simplifying assumption regarding (2)—that the functional relationships between elements in the system are linear. Then we can write,



(3)

where the a's are the causal regression coefficients of the system. One of these can be estimated directly from the data,

$$\hat{a}_2 = b_2 \qquad (4)$$

where b_2 is the observed linear regression of Hunters on Moon. Unfortunately, there is only one more independent regression coefficient which can be estimated from the observations. This is b_3, the observed regression of Kill on Moon. Now b_3 can be equated to the estimated causal coefficients,

$$b_3 = \hat{a}_3 + \hat{a}_1\hat{a}_2 \quad , \qquad (5)$$

which says that the effect of Moon on Kill is the sum of the direct effect and the indirect effect through Hunters. Substituting (4) into (5) we obtain

$$b_3 = \hat{a}_3 + \hat{a}_1 b_2 \quad . \qquad (6)$$

Thus, we have one equation in two unknowns, and cannot solve for $\hat{a}_1$ and $\hat{a}_3$. (There are, in fact, an infinite number of solutions to this equation. For any arbitrary value of $\hat{a}_3$, there is a value $\hat{a}_1$ which satisfies the equation, and vice versa.) Again, you may be wondering just what is the purpose of all this, if after building this foundation, we still cannot estimate the causal regression coefficients in which we are interested. It was pointed out earlier that to ignore a part of such a system, and analyze part of it, was to stand the chance of making gross errors. This is obvious from consideration of (3). Suppose that a_3 was actually zero—that there was no direct effect of Moon on Kill, but that hunters thought there was such an effect, and hunted accordingly. Then one could find a significant regression of Kill on Moon, even though there was no causal effect. From (5), we see that in such a case, this regression coefficient would really be an estimate of the product of a_1 and a_2. Thus, the use of the path diagram aids in correct interpretation of the data. It also aids in detecting incorrect interpretation!

Another value of the path diagram is that it provides clues as to possible ways in which to increase the information available from the system. For example, we might be willing to specify that a_1 is greater than zero and smaller than 1, so that, from (6), we can write

$$0 < a_1 < 1$$

so that, from (6), we can write

$$0 < \frac{b_3 - a_3}{b_2} < 1$$

$$(b_3 - b_2) < \hat{a}_3 < b_3 \, . \qquad (7)$$

That is, by making assumptions about some parts of the system, one can evaluate other parts of the system. Similarly, it can be shown that when parts of the system are ignored in the analysis, one is implicitly making restricting assumptions. Furthermore, the path diagram can aid in specifying the assumptions necessary to any possible conclusion.

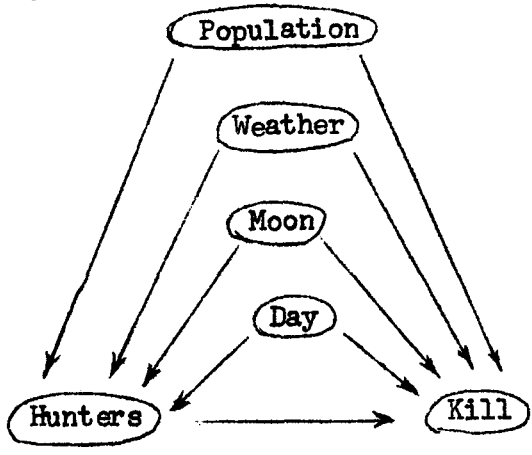
These comments have skirted an idea which is perhaps the most important of all. If it is possible to draw conclusions after assuming, say, a_2 to be zero, why not *make* a_2 zero so that the conclusion is not dependent on the assumption. This is the experimental approach: one can make a_2 zero by the simple expedient of randomizing hunters over days. Then the partial regression coefficients of Kill on Hunters and Kill on Moon are estimates of a_1 and a_3. In fact, one can easily go a step further in this case, tabulate the moon phases before the hunting season, and work out a balanced experimental design between Moon and Hunters. This randomization also serves as insurance against the presence of unspecified components of the system, which is the reason that experimental people seldom are concerned about path analysis.

The problem is that very seldom can randomization be accomplished with the sort of variables that we work with. However, although it would certainly be impossible to break the causal path between Moon and Hunters in the present example, (these are private lands) it might be possible to bend it a bit!

Too, one may sometimes be able to incorporate a variable that hypothetically affects Hunters but not Kill. There are a number of ways to improve the interpretation of observational data that will become evident from the study of the path diagram.

THE GOOSE DATA

To return to the analysis of the goose kill data, we begin with the following diagram:



(8)

Note that no causal paths connect the co-variates, day, moon, weather and population. There will be correlations between these, but in the present example, we have assumed no causal relationships. Note, too, that we have not indicated a feedback from hunters to population. On a seasonal basis, this would perhaps be operative, but we have set the analytic unit to be a day, and there is not time within a day for the goose population to adjust to the number of hunters.

Some discussion of the variables is in order. Some, such as kill, hunters, and population, require no coding, while others must be assigned arbitrary quantitative values. This is a crucial step in the procedure, and actually is a part of the model building. In the present case, we were more successful in quantifying some of the variables than others. Following are the variables in order they were treated in the analysis.

0. *Day.* This is the chronological day of hunting, 1 through 52.
1. *Population level.* Estimates of the goose population on Lake Mattamuskeet were available at approximately two-week intervals. Linear interpolation was used to obtain estimates for hunt days between the counts.
2, 3. *Moon: linear, quadratic.* Each day was designated as following a "light," "intermediate," or "dark" night, as evidenced by the moon stage. It is difficult to specify the rules on this, but in general it was attempted to separate each "moon month" into three classes of the same number of days. In order to better understand the effect of moon, two variables were defined, linear with codes -1, 0, 1 for dark, intermediate and light, and quadratic with codes 1, -2, 1 for the same phases.
4. *Cloud.* The sky was judged to be either cloudy (1) or not cloudy (0) each morning at 8:00 o'clock. This is not a very satisfactory characterization of this variable.
5. *Wind direction.* All northerly (NW, N, NE) winds were coded (1), and all others (0), the observation made at 8:00 A. M. This is also not a very satisfactory representation of wind direction.
6. *Wind velocity.* Coded as 0, 1, 2 indicating 0-10, 10-20, and 20+ m. p. h., estimated at 8:00 A. M.
7. *Temp min.* At 4:00 P. M., the minimum temperature during the preceding 24 hours was recorded.

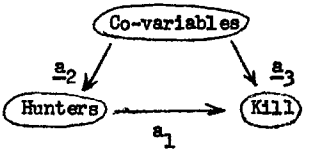
8. *Temp. max.* At 4:00 P. M. the maximum temperature during the preceding 24 hours was recorded.
9. *Rainfall.* Rainfall was coded as 1 or 0, depending on whether it rained during the previous 24 hours. Records taken at 4:00 P. M.
10. *Hunters.* The number of hunters utilizing the reporting fields on the day in question.
11. *Goose kill.* The reported number of geese killed on the reporting fields on the day in question.

ANALYSIS

The regression analysis of the path diagram illustrated in Figure (8) requires the computation of two sets of partial regression coefficients, b_2 and b_3 *where* these vectors are defined as follows:

$$\underline{b}_2 = \begin{bmatrix} b_{20} \\ b_{21} \\ \cdot \\ \cdot \\ \cdot \\ b_{29} \end{bmatrix} \quad \text{and } \underline{b}_3 = \begin{bmatrix} b_{30} \\ b_{31} \\ \cdot \\ \cdot \\ \cdot \\ b_{39} \end{bmatrix} \tag{9}$$

and where b_{20} is the partial regression coefficient of Hunters on Day, b_{21} the partial regression coefficient of Hunters on Population Level, b_{30} the partial regression coefficient of Kill on Day, b_{31} the partial regression coefficient of Kill on Population Level, etc. In each case, the partial regression coefficients are with respect to the particular variable, with all others (excluding Hunters and Kill) held constant. Using this vector notation as a sort of shorthand, we can diagram the system (8) as follows:



(10)

Then we can write, from analogy to (3),

$$\begin{aligned} \hat{\underline{a}}_2 &= \underline{b}_2 \\ \underline{b}_3 &= \hat{\underline{a}}_3 + \hat{a}_1\hat{\underline{a}}_2 \\ \text{so that} \quad \underline{b}_3 &= \hat{\underline{a}}_3 + \hat{a}_1\underline{b}_2 \ . \end{aligned} \tag{11}$$

The two sets of partial regression coefficients were computed on the 650 digital computer, so that we have from (11) the following 10 linear equations in 11 unknowns (presented in vector form):

$$\hat{\underline{a}}_3 = \begin{bmatrix} -5.91631 \\ -1.34046 \\ -19.79421 \\ -9.65096 \\ 38.08798 \\ 11.59054 \\ -.72628 \\ .94177 \\ .59655 \\ 21.44332 \end{bmatrix} - \hat{a}_1 \begin{bmatrix} -5.74412 \\ -1.10857 \\ -20.09145 \\ -17.11605 \\ -10.77163 \\ .80694 \\ -15.01324 \\ 1.20749 \\ -.50554 \\ 30.71949 \end{bmatrix} \qquad (12)$$

As before, when we had one equation with two unknowns, there are an infinite number of solutions to this set of equations, and we still have not obtained the answers we want. Several things occur to us, however, that might shed some light on the problem. First, it is evident from the data that there is still a significant amount of variation in the residuals of Kill (after fitting to variables 0-9) that is accounted for by the residual variation in Hunters. We can estimate a regression coefficient, then, from these residuals. We designate by b_1 the estimated linear regression of Kill on Hunters, with variables 0-9 held constant. In the present data,

$$b_1 = .67857 \ .$$

Substituting $b_1 = \hat{a}_1$ into (12), we obtain

$$\hat{\underline{a}}_3^{(1)} = \begin{bmatrix} -2.01852 \\ -.58822 \\ -6.16075 \\ 1.96348 \\ 45.39728 \\ 11.04297 \\ 9.46125 \\ .12240 \\ .93959 \\ .59800 \end{bmatrix} \qquad (13)$$

Incidentally, this solution for $\hat{\underline{a}}_3$ and $\hat{a}_1$ is the same that would be gotten by the straight multiple regression of Kill on variables 0-10. What then is the advantage of the present approach? We have clearly specified the assumptions necessary to this solution for $\hat{a}_1$ (and $\hat{a}_3$) to wit: there are unknown (or rather unspecified) forces causing variation in Hunters, but exerting no influence on Kill, either directly or through variables 0-9.

Earlier, we commented that if one of the coefficients in $\hat{a}_3$ was actually zero (and we knew it was zero), then it would be possible to solve the equation (12). Suppose, for example, that we had evidence (or knew from theory) that $a_{31}=0$. Then b_{31} would be an estimate of a_1a_{21}, and we already have an estimate of $a_{21}=b_{21}$. Hence, $\hat{a}_1=b_{31}/b^2_1$. Solution to the remaining components of $\hat{a}_3$ follows directly. In our example, we could, by way of illustration, assume that a_{37} was zero. Then, solving (12), we obtain, and

$$0 = .94177 - \hat{a}_1(1.20749)$$

$$\hat{a}_1 = \frac{.94177}{1.20749} = .77994$$

and

$$\underset{\sim}{a}^{(2)}_3 = \begin{bmatrix} -1.43624 \\ -0.47584 \\ -4.12408 \\ 3.69853 \\ 46.48921 \\ 10.96118 \\ 10.98315 \\ 0.00000 \\ 0.99084 \\ -2.51604 \end{bmatrix} \qquad (14)$$

It would appear that $\hat{a}_3(^2)$ tends to corroborate $\hat{a}_3$ $(^1)$, but this is not the case, as we selected a_{37} as a likely candidate to assume to be zero on the basis of $\hat{a}_{37}$ $(^1)=.12240$. Therefore, the assumptions, and consequences, of the two solutions are much the same. What we really need is some variable which, a priori, we could feel sure had an effect on hunting pressure but none directly on Kill or the other variables in the model. Such a variable is difficult to imagine, but may conceivably exist, or be manufactured as an experimental element in the observational system.

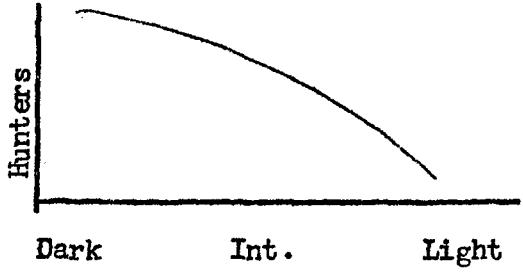
In any event, we have no firm basis on which to estimate causal regression coefficients in the present analysis of the Mattamuskeet goose field hunting system. However, we have gained a good deal more information about the system that we would have had we not conducted the analysis. Some examination of some of these things is in order. We will restrict this examination to the first set of solutions, purely for convenience.

INTERPRETATION OF THE GOOSE DATA

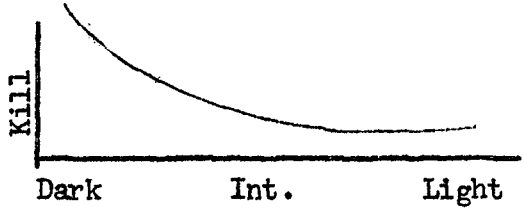
Our first estimate of a_1, $b_1(^1)=.67857$, indicates an increase of .67857 geese killed for each additional hunter, when all other variables are held

constant. This is a bit disconcerting as the average kill per hunter for the season was 5237/7588=.69017. Have we gone to this detail to modify the estimate of mean kill so little? The one, $b_1(1)$, is an estimate of this rate of change with a number of variables (conditions) held constant, while the other, $\bar{y}$, is an estimate of this rate of change averaged over these conditions as they chanced to exist during the one season. They are not conceptually the same, even if they are numerically almost the same!

The effect of moon is very interesting! It is seen from examination of b_3 that both the linear and quadratic coefficients of hunters on moon are negative. That is, the curve of hunters on moon is of the following form:



Then, note from $\hat{a}_3(1)$ that the linear coefficient of kill on moon is negative, but the quadratic coefficient is positive! That is, the relationship between kill and moon is something as follows:



Thus, we have evidence that although the hunter correctly interprets the dark of the moon as the best time to hunt, he has a tendency to associate the intermediate phases with the dark phases. On the other hand, the goose associates intermediate moon phases with light phases. This result is not surprising, as only an hour or two of moonlight is required in order for the geese to feed, and this is unavailable only on a very few days during the moon month.

In several other variables, the hunters' response was quite different from the goose's response. For example, cloud cover at 8:00 caused a large increase in hunting success, but a noticeable decrease in hunting pressure. Wind direction had little influence on pressure, but an appreciable effect on success. Wind velocity demonstrated a negative effect on pressure and a positive effect on success. Rain during the 24-hour period seemed to greatly increase hunting pressure, but have little effect on the hunting success. The real significance of these observations is not yet clear, and a further look at more data is indicated.

It was expected that hunting pressure would drop off as a function of time, and this is substantiated by the analysis. Note that success, too decreases during the season, other variables being held constant. This indicates either a learning process on the part of the geese, or a selection process (selection for more wary geese or less capable hunters) or both. It will be interesting to follow this through several more years.

SUMMARY

The distinction is made between observational data and experimental data, and the analytic and interpretive consequences discussed. The method of path analysis is presented as an aid in this analysis and interpretation. Illustration is made by an analysis of kill data from goose fields around Lake Mattamuskeet, North Carolina in the 1960-61 season. The method proves helpful in defining the system, and several interesting interpretations are made.

BIBLIOGRAPHY

1. Mallios, W. S. 1961. Some Aspects of Linear Regression Systems. Inst. of Stat. Mimeo Series No. 298, N. C. State College, Dept. of Exp. Stat., Raleigh, N. C., 121 pp.
2. Tukey, J. W. 1954. Causation, Regression and Path Analysis. Chapter 3. Statistics and Mathematics in Biology, ed. Kempthorne, Bancroft, Gowen and Lush, Ames, Iowa: The Iowa State College Press.
3. Turner, M. E. and C. D. Stevens. 1959. The Regression Analysis of Causal Paths. Biometrics 15(2) : 236-258.
4. Wright, S. 1934. The Method of Path Coefficients. Annals of Math. Stat. 5: 161-215.
5. Wright, S. 1960. The Treatment of Reciprocal Interaction, With or Without Lag, in Path Analysis. Biometrics 16(3) : 423-445.

RESULTS OF DESIGN TESTS OF METHODS OF ESTIMATING DOVE HARVEST *

By HERBERT STERN, JR.,
Louisiana Wild Life and Fisheries Commission

By W. SCOTT OVERTON,
North Carolina State College

By LAWRENCE SOILEAU,
Louisiana Wild Life and Fisheries Commission

By EUGENE LEGLER, JR.,
Tennessee Game and Fish Commission

I. INTRODUCTION

During the 1960 hunting season, Louisiana and Tennessee conducted a pilot study to determine the feasibility of using the telephone and field sampling frames to estimate hunter kill of mourning doves. This study was requested by the Dove Committee of the Southeast Section of the Wildlife Society after theoretical sampling methods were explored and reported on by Chapman, Overton and Finkner (1959).

Methodology and a cursory inspection of data obtained from the pilot study was reported by Legler, Stern and Overton (1961) and the reader should refer to that publication for details regarding operational procedures. In the present paper are presented analytic and estimation procedures, and an evaluation, of the general method from the standpoint of the data collected.

The survey was based on a complex frame, consisting of two sub-frames, which was considered by Chapman *et al.* (1959) to be theoretically the most promising of all frames studied. This complex frame consisted of:

1. The primary frame of telephone subscribers. In our field test we used two exchanges in Louisiana and one in Tennessee. From each of these ex-

* A contribution of the Tennessee Game and Fish Commission, the Louisiana Wild Life and Fisheries Commission, through Federal Aid to Fish and Wildlife Restoration Project FW-2R, and from the Southeastern Cooperative Fish and Game Statistics Project, Institute of Statistics, North Carolina State College.