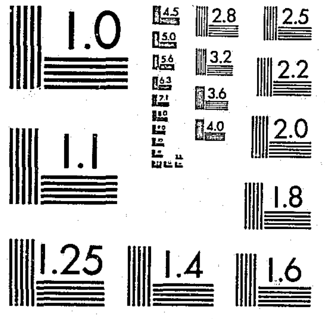


MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

9496

LONGITUDINAL MODELS, MISSING DATA, AND THE ESTIMATION OF VICTIMIZATION PREVALENCE

by

William F. Eddy, Stephen E. Fienberg,

and Diane L. Griffin

DEPARTMENT
OF
STATISTICS

89245

Carnegie-Mellon University
PITTSBURGH, PENNSYLVANIA 15213

LONGITUDINAL MODELS, MISSING DATA, AND THE ESTIMATION OF VICTIMIZATION PREVALENCE

by

William F. Eddy, Stephen E. Fienberg,
and Diane L. Griffin

Technical Report No. 231
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

January, 1982

7

1. Introduction

The National Crime Survey (NCS), designed and executed by the U.S. Bureau of the Census, produces on an ongoing basis, national data on crime victimization in the U.S.A. Based on a stratified multistage cluster sampling plan, the NCS utilizes a rotating panel of household locations. We give further details of the sample design in Section 2 (see also Fienberg 1978, 1980b). From its inception in 1972, the NCS has been used almost exclusively to produce incidence rates by type of crime and selected characteristics of the victims and/or offenders, of the sort found in NCS annual reports (e.g. see U.S. Department of Justice, 1981b).

In March 1981, the Bureau of Justice Statistics (BJS), which sponsors the NCS, issued a report (U.S. Department of Justice, 1981a) on the prevalence of crime, in which the key quantity estimated was the percentage of households touched by crime in a given year. In this paper, we describe some stochastic longitudinal models for victimization that can be used to produce such an annual prevalence rate, and we show how the BJS's reported prevalence measure relates to those that we have derived. Moreover, following a suggestion of Albert Biderman, we adopt a cheery approach to the otherwise depressing prevalence measure by taking its complement -- the percentage of crime-free households in a given year.

For purposes of this paper we need to distinguish amongst the housing unit (HU) or location, the household (HH) or family living in that unit, and the individuals who compose the household. As we note in Section 2, at each interview NCS respondents provide victimization information on the preceding 6 months. To actually determine whether a HU (or a HH) has been "touched by crime" it is, in principle, necessary to examine all of the interviews of the occupants of the HU (or members of the HH) that contain information for some part of the year in question. Typically this will mean that we need information from a respondent for 3 successive interviews to reconstruct the victimization profile for a single year, and that the data will need to be matched or linked in some type of longitudinal format.

In practice, we do not get to see a complete longitudinal record for each housing unit, household, or individual for any specific 12-month period of interest. When a HU enters or leaves the sample during the year, part of the desired data will be missing. Similarly data for 6-month intervals can be missing for individuals or households (HH) due to non-interviews. Finally, if the NCS is viewed as an HH sample rather than an HU sample, then missing data can occur as a result of households and individuals who move between interviews. Any attempt at constructing prevalence indicators of crime must directly address the problems of missing data, and their relationship to the data that are not missing. We give some clues as to the dimensions of the missing data problem for our prevalence measures in Section 4, linking actual missing data rates to officially published nonresponse rates.

From a methodological research perspective our interest would normally focus on the development of stochastic models for longitudinal victimization records (see e.g. the discussion in Fienberg 1978, 1980b), and such a perspective remains a critical feature of the cheery indicator problem. Some simplifications ensue when we restrict our attention to prevalence measures, however, and these allow us to make progress on a modelling problem that would otherwise appear to be virtually intractable. For example, longitudinal modelling typically would require a time-ordered victimization history, but, as Reiss (1980) and Fienberg (1980b) note, NCS data have ordering problems when series victimizations or multiple victimizations occur. From a prevalence perspective, these ordering problems do not really matter -- all we need to know is that at least one victimization (perhaps of a given type) has occurred. Such simplifications, when combined with the interest in prevalence measures by the Bureau of Justice Statistics, has guided our methodological efforts.

In this paper we develop several "naive" stochastic longitudinal models in which missing data are assumed to be missing at random. Fragmentary evidence available from analyses by Reiss and others suggests little support for such an assumption, and efforts to model "missingness" will be part of our future activities. We refer to the models described here as "naive" because each is based on a large number of inappropriate but simplifying assumptions, and because they reflect little of the structure described in longitudinal analyses by Reiss (1980) and by Fienberg (1980a, 1980b), for example. Moreover, we fit the models in Section 6 only to data on HU's, not to longitudinal files on HH's or on individuals, and we treat victimizations in an aggregate form, not distinguishing among types. We present these "naive" models and results from their preliminary application to establish a starting point for future modelling and analysis efforts that we hope will incorporate more appropriate and substantively interesting assumptions.

2. NCS Sample Design

In this Section, we give a brief synopsis of the rotating panel design of the NCS, because this structure is so critical to an understanding of the NCS longitudinal data base discussed in Section 3. For further details see Fienberg (1980b) and U.S. Department of Justice (1980, 1981b).

The NCS is based upon a stratified multistage cluster sample. The first stage consists of dividing the United States into 1931 primary sampling units (PSUs) comprising counties or groups of continguous counties. The PSUs are then separated into 376 strata, 156 of which are self-representing. From the remaining 220 strata one PSU is selected from each stratum with probability proportional to population size. Within each of the 376 PSUs selected, a systematically chosen group of enumeration districts is selected, and then clusters of approximately four HU's each are chosen within each enumeration district. This method produces a self-weighting probability sample of dwelling units and group quarters within each chosen PSU. For 1979, this process led to the designation of about 62,000 HU's, and interviews were obtained from occupants of about 51,000. Most of the remaining designated HU's were vacant or otherwise deemed to be ineligible for inclusion in the NCS; about 2,200 of these HU's would actually be labelled as non-respondents.

The basic sample is divided into six subsamples or rotation groups of about 9,000 HU's each. The rotation groups are numbered from 1 through 6 within each sample. Every six months a new rotation group enters the sample and the "oldest" existing rotation group from the previous sample is dropped. Each rotation group is divided into six panels with panel 1 being interviewed in January and July, panel 2 in February and August, etc. This process spreads the workload of the field staff. Each HU is in the survey for three full years for a total of seven interviews. (There are, however, some HU's that, due to the initial implementation of the rotation schedule, actually had as many as eight or nine interviews). The data collected at the first of the seven interviews are used for bounding purposes, i.e., to establish a time frame intended to avoid duplication of victimization information in subsequent interviews. These data are not incorporated into the official BJS reported rates (either incidence or prevalence) but have been incorporated in rates reported in this paper.

Table I shows the rotation scheme. For instance, in September, panel 3 in each of the rotation groups 2 through 6 of sample A and 1 and 2 of sample B will be interviewed. In the following March, panel 3 of rotation group 2 in sample A is replaced by panel 3 of rotation group 3 in sample B.

TABLE I
Rotation Scheme
(The numerical entries represent panel numbers within samples)

	Rotation Group											
	Sample A						Sample B					
Month	1	2	3	4	5	6	1	2	3	4	5	6
Jan	1	1	1	1	1	1	1					
Feb	2	2	2	2	2	2	2					
Mar	3	3	3	3	3	3	3					
Apr	4	4	4	4	4	4	4					
May	5	5	5	5	5	5	5					
June	6	6	6	6	6	6	6					
July		1	1	1	1	1	1	1				
Aug		2	2	2	2	2	2	2				
Sept		3	3	3	3	3	3	3				
Oct		4	4	4	4	4	4	4				
Nov		5	5	5	5	5	5	5				
Dec		6	6	6	6	6	6	6				
Jan			1	1	1	1	1	1	1			
Feb			2	2	2	2	2	2	2			
Mar			3	3	3	3	3	3	3			
Apr			4	4	4	4	4	4	4			
					⋮							

At each interview information is acquired on the "household" as well as on all persons age 12 or older living in the designated HU. The interview questionnaires are used to record information about the household and the persons comprising it, as well as details on victimization events occuring during the previous 6 months. There is no guarantee of continuity of households or persons in the sample. If a household moves and is replaced by a new one, the experiences of the new household and its members are recorded at the time of the next interview. If the household composition changes, only information on those persons who are currently members is recorded. As a consequence, despite the fact that the entire rotation group of housing units provides "bounding information" at each interview, there is no bounding information available for a large proportion of households and especially persons for any given interview.

NCS data are aggregated on a quarterly basis to produce quarterly estimates of the volume and rates of victimization. Annual estimates are produced by pooling quarterly estimates. Care must be taken to distinguish *collection* months (i.e., the month in which data are collected) from *reference* months (i.e., the month to which the data relate). Data are actually stored by collection quarter (3 months), each of which contains data for 8 reference months. Conversely, sample data from 8 collection months are required to produce estimates for each reference quarter. For a full reference year, data from 17 collection months are used, involving 8 rotation groups and 47 panels. More detailed discussions of the relation between reference and collection months may be found in Fienberg (1980b) and Penick and Owens (1976).

The NCS questionnaire distinguishes between individual identifiable incidents, and series of at least three similar incidents which the respondent is unable to separate in time and place of occurrence. Either may be personal or household victimizations. For individual victimizations, the questionnaire records the month in which the crime took place. For series victimizations, the method of recording involves the details for only the most recent event in the series, and the date of *first* occurrence. Prior to 1979, the respondents were asked to indicate the number of incidents (3–4, 5–10, 11+) and the quarter(s) in which the incidents took place, i.e., Winter (December to February), Spring (March to May), Summer (June to August), and Fall (September to November). In January 1979 this procedure was altered, and now respondents provide the number of incidents (not necessarily using the earlier grouping), and a breakdown of this count into quarters of the year (January to March, April to June, July to September, and October to December), rather than simple indicators for seasons.

In this paper we include series victimizations, but we treat them as having occurred only in the first month in which the series occurred. This use of series victimizations reflects the way in which the data were coded for 1972–1976, the time frame of our longitudinal files, and not how they should be used as a consequence of the more detailed reporting scheme initiated in 1979. The estimates clearly lead to underestimates of the extent of victimization.

3. NCS Data Base

The NCS victimization data are publicly available through the Inter University Consortium for Political and Social Research (ICPSR) at the University of Michigan. These data are grouped into quarterly collection files which include records of all the interviews completed by the U.S. Bureau of the Census for a particular three–month period. Because the occupants of a specific housing unit are interviewed every six months, each quarterly collection file contains the records for at most one interview for that housing unit.

There are three types of information collected by the NCS: Household Items, Individual Items, and Crime Incident Reports. Because households have varying numbers of individuals, and individuals report varying numbers of crime incidents it is not sensible to think of these data (nor is it feasible to store them) in a rectangular array with each row representing a housing unit, household, or even individual. ICPSR stores the data in an OSIRIS IV hierarchical file. OSIRIS IV is a proprietary software package developed and maintained at the Survey Research Center at the University of Michigan. The OSIRIS IV file is a sequential file which can also be interpreted as a three–level hierarchy: household (H), person (P), and incident (C). The sequential order of the data records collected at one interview for two particular households, the individuals within those households, and the crime incidents reported by those individuals might be as follows:

Household record for HU #1
Person record for P #1 in HU #1
Incident record for C #1 for P #1 in HU #1
Person record for P #2 in HU #1
Incident record for C #1 for P #2 in HU #1
Incident record for C #2 for P #2 in HU #1
Person record for P #3 in HU #1
Household record for HU #2
Person record for P #1 in HU #2
Person record for P #2 in HU #2
Incident record for C #1 for P #2 in HU #2
etc.

The first household has three individuals; the first individual reported one crime incident and the second individual reported two crime incidents. The second household has two individuals; the second individual reported one crime incident.

Six months later (two collection quarters) these two housing units might be eligible for another interview. In that case the relevant quarterly OSIRIS file would contain similar data records for the two housing units. Of course, the exact pattern might be different, as the occupants of the housing unit may change and the reported incidents, if any, will be different. It has been convenient for the analyses, which we describe below, for us to have these data reorganized into longitudinal files with all of the data for one housing unit together in chronological order, rather than scattered over many quarterly collection files. While we have not yet completed this reorganization directly from the ICPSR tapes, we were fortunate to obtain, from Professor Albert Reiss of Yale University, longitudinal files covering the period from July 1, 1972 to December 31, 1976. These longitudinal data are divided into three separate files: the first contains only the household records; the second contains only the person records; and the third contains only the crime incident records. Each one of the three files can be regarded as a rectangular array with each row being respectively a household interview record, a person interview record, or a crime incident report.

The structure of the three longitudinal files for the same two particular housing units described above is as follows, assuming the first housing unit was interviewed for the third time and the second housing unit was interviewed for the first time during the collection quarter indicated above. The information from above is indicated by italics.

Household File

Household records for HU #1 at first 2 interviews

Household record for HU #1 (Third Interview)

Household records for HU #1 at subsequent interviews

Household record for HU #2 (First Interview)

Household records for HU #2 at subsequent interviews

etc.

Person File

Person records for P #1 in HU #1 at first 2 interviews

Person record for P #1 in HU #1 (Third Interview)

Person records for P #1 in HU #1 at subsequent interviews

Person records for P #2 in HU #1 at first 2 interviews

Person record for P #2 in HU #1 (Third Interview)

Person records for P #2 in HU #1 at subsequent interviews

Person records for P #3 in HU #1 at first 2 interviews

Person record for P #3 in HU #1 (Third Interview)

Person records for P #3 in HU #1 at subsequent interviews

Person record for P #1 in HU #2 (First Interview)

Person records for P #1 in HU #2 at subsequent interviews

Person record for P #2 in HU #2 (First Interview)

Person records for P #2 in HU #2 at subsequent interviews

etc.

Incident File

Incident records for P #1 in HU #1 at first 2 interviews

Incident record for C #1 for P #1 in HU #1 (Third Interview)

Incident records for P #1 in HU #1 at subsequent interviews

Incident records for P #2 in HU #1 at first 2 interviews

Incident record for C #1 for P #2 in HU #1 (Third Interview)

Incident records for P #2 in HU #1 at subsequent interviews

Incident records for P #3 in HU #1 at first 2 interviews

Incident records for P #3 in HU #1 - None at Third Interview

Incident records for P #3 in HU #1 at subsequent interviews

Incident records for P #1 in HU #2 - None at First Interview

Incident records for P #1 in HU #2 at subsequent interviews

Incident record for C #1 for P #2 in HU #2 (First Interview)

Incident records for P #2 in HU #2 at subsequent interviews

etc.

In the Household File there is one record for each interview. Thus, the lines not in italics in this example represent data records. In the Person File and in the Incident File there are records only when data are actually collected. As a consequence, the lines not in italics in this example may or may not represent actual data records. This example does not reflect the full complexity of the Person File and the Incident File because the individuals within a household may vary from interview to interview and because the household itself (within the housing unit) may change from interview to interview. We have ignored this additional complexity in the analyses described below, by focussing on housing units and not on households.

Data for the housing units in each of the three files are stored in the same order with respect to an internal identification number. This makes cross-references among the three files fairly simple. Although the data can still be thought of in the household-person-incident hierarchy, it is simpler to conceptualize questions and execute programs on rectangular files. When more than one level of the hierarchy is involved in an analysis special programming will be required (just as with the OSIRIS IV hierarchical files) but since each of the three files is rectangular the job will be simpler.

For the purposes of this paper we have chosen to work, not with the full longitudinal data base, but rather with a systematic random sample of 1539 household locations (every 100th HU). As a consequence we need not really address ourselves to the current controversy regarding the use of sample weights in model-based statistical analyses (for some discussion of this matter see Fienberg, 1980b). Virtually all clustering effects are removed as a result of the systematic sampling and the estimates described in Section 5 are derived for simple random sampling (i.e. ignoring the NCS sample design). All of the results reported in Section 6 are computed from this sample of 1539 HU's and thus are subject to substantially greater variability than those estimates that we ultimately plan to compute for the full NCS data base.

4. Missing Data and Weighted Analyses

One of the most troublesome aspects of longitudinal modelling for sample survey data is the handling of missing data. In a cross-sectional analysis aggregate quantities are typically estimated by applying assigned weights to sample units. As an example, in the NCS for personal crimes these weights are the product of a "basic weight" times a "within household noninterview factor" times a "duplication control" times a "household noninterview factor" times a "first-stage ratio-estimate factor" times a "second-stage ratio-estimate factor." The two

"noninterview factors" are adjustments for missing data that are computed, within appropriate subgroups, as the ratio of the count of possible interviews to the count of actual interviews.

For cross-sectional analysis missing data involving full records for a given month is a problem of limited magnitude. For longitudinal analyses it remains to determine the magnitude of the missing data problem. To this end we counted, for each of the 1539 HU's in our sample, the number of months of missing data. Because of the fact that interviews cover 6-month periods Table II below is organized with 6 months in each row. For example, the entry in the row labelled 2 and the column labelled 3 gives the number of HU's with (6×2)+3 = 15 missing months. We note that all 1539 HU's are included in Table II regardless of the length of time between July 1972 and December 1976 that they were listed as housing units participating in the survey. For instance, some of the HU's in the (0,0) cell were in the survey for only 1 interview while some were in the survey for 8. Moreover, a household with only 1 interview may contribute as little as 1 month to the reference period of interest.

TABLE II

Number of Housing Units in Sample
with 6K + I Months of Missing Data

K \ I	0	1	2	3	4	5
0	916	8	7	10	15	5
1	229	5	3	3	2	5
2	100	2	3	1	4	2
3	65	1	5	1	1	1
4	49	2	5	4	0	2
5	25	1	0	2	3	2
6	17	3	2	1	0	1
7	12	2	0	0	0	3
8	6	1	0	0	0	2

In our sample of 1539 HU's, only 916, or about 60%, have complete records for that part of the July 1972 to December 1976 time-period. According to the most recent NCS report (U.S. Department of Justice 1981b), 96% of all eligible HU's participated in the survey. On the average a HU from our 1% sample for 1972–1976 was in the survey for 4.25 interviews. Thus, if we assume that missing interviews occur independently of one another, we would expect roughly $(0.96)^{4.25}$ or about 84% of our HU's to have complete records for the July 1972 to December 1976 time period. The 96% nonresponse figure is not really applicable here, however.

BJS reports that for 1979, of the approximately 62,000 HU's sampled, interviews were obtained from the occupants of about 51,000. Of the remaining 11,000 HU's, about 2,200 were occupied by persons who were not interviewed because they could not be reached and about 8,800 were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible. The 96% figure is the percentage of HU's deemed eligible that responded, that is, 51,000/(51,000+2,200) = 96% . Table II includes *all* sampled HU's and thus the appropriate response figure is 51,000/62,000 or about 82% . Thus, if we assume as before that missing interviews occur independently of one another, we would expect roughly $(.82)^{4.25}$ or 43% of the HU's in the 1% sample to have complete records, substantially less than the observed 60%.

The U.S. Bureau of the Census (undated) reports that about 2% of the processed HU's are units that had no address listed on the listing sheet, had been demolished, had moved (i.e. trailer), had been converted to a business, or had merged with another unit. Removing these units from the 62,000 sampled units yields a response rate of about 84% . With this response rate, assuming again that missing interviews occur independently of one another, we would expect roughly 48% of the HU's in the 1% sample to have complete records. The fact that about 60% of the 1539 HU's have complete records suggests that the missing interviews may not occur independently of each other and that missingness may be positively correlated over time.

Even if the proportion of records with missing data were not as great as that indicated by Table II, we believe that it still would not be advisable to construct weighted aggregates using the weights described at the beginning of this section, and to do "weighted analyses" of the longitudinal data base. This is simply because the weights change from interview to interview! It is tempting to argue that the weights should not change much and one could use an "average" weight. Such is not the case. Several households in our sample of 1539 exhibited substantial variation in the sample-based weights. For example, in one selected unit, the household weight varied from 964.996 to 1120.389, while the weight for one of the persons in the household varied from 874.662 to 1389.469. We believe that a more thoughtful model-based approach, rather than the blind application of sample-based weights, is required to take into account the implication of the sample design on statistical analysis (see the related discussion in Fienberg, 1980b).

5. Models for Victimization Prevalence Measures

The analysis of data sets from which some items are missing is often performed by ignoring the missing components. Such an analysis will lead to appropriate inferences only if the nonresponse mechanism does not depend on the values of the missing items. In particular, the missing data may be ignored if the assumption is made that whether or not an HU responds at a particular interview does not depend on whether or not that HU was victimized in the six months prior to that interview. When this condition holds we will say that the missing data are missing at random (Little, 1980; Rubin, 1976). Although the validity of the missing at random condition cannot be checked directly, some preliminary analyses show that HU's with high proportions of missing data tend to exhibit higher rates of victimization than HU's with low proportions of missing data. The full impact of the missing at random assumption still needs to be assessed.

The remainder of this section is devoted to the exposition of several estimators of the percentage of crime-free HU's, θ. Some of these estimators are based on specific models of victimization and the missing at random assumption and others are more ad hoc in nature.

Estimator 1: An Ad Hoc Approach

We begin by considering an ad hoc estimator of the proportion of HU's not victimized by crime in a year:

1) Consider each HU that had an interview covering at least one month of the year.

2) If there is no victimization reported as having occurred in the year, consider the HU as being crime-free.

3) For each HU determine the number of interview months in the year (that is, the number of months covered by an interview).

4) Use an an estimator

$$\hat{\theta}_1 = \frac{\text{number of interview months for crime-free HU's}}{\text{total number of interview months}} . \qquad (1)$$

If a HU has reported a victimization in any month of the given year then we are sure that it was not crime-free. With this in mind we may prefer a variation of $\hat{\theta}_1$:

$$\hat{\theta}'_1 = \frac{\text{CFM}}{12 \times \text{number of victimized HU's} + \text{CFM}} \qquad (2)$$

where CFM is the number of interview months for crime-free HU's (that is, the numerator of $\hat{\theta}_1$). Note that $\hat{\theta}'_1 \leq \hat{\theta}_1$.

Both $\hat{\theta}_1$ and $\hat{\theta}'_1$ have built-in biases that we expect will lead them to be overestimates (perhaps by a substantial amount) of some true proportion, θ. We note that they are, at least in spirit, similar to estimators, such as the Kaplan-Meier estimator, that appear in the survival analysis literature (e.g. see Kalbfleisch and Prentice, 1978).

Estimator 2: The BJS Estimator

Alexander (1981) describes two different estimators which were used to produce the prevalence rates published by the BJS (U.S. Department of Justice, 1981a). Because BJS needed to estimate rates for 1980 before the end of March, 1981, and because it would be necessary to have information from interviews through June, 1981 in order to calculate 1980 rates, BJS used information from 1980 interviews in place of information from interviews that would have occurred after January 1981. Thus, the BJS estimators used only 2 interview records rather than the 3 that we require to obtain the information for an HU for a full calendar year. For example, BJS used the March 1980 and Sept. 1980 interviews for a HU in panel 3 although the information obtained at these interviews is actually for the year Sept. 1979 through Aug. 1980 (see Section 2).

In Appendix I we present a detailed description of the BJS estimates. Both are of a form similar to $\hat{\theta}_1$, except for the following:

(a) the time periods for interviews contributing to a given annual estimate extend backwards out of the period of interest,

(b) different weights are applied to different types of non-interviews,

(c) $1-\theta$, that is, the probability of being victimized in a year, is estimated rather than θ.

In both cases the denominator of the estimator is interpretable as the number of interview months (divided by 12). For the first estimator, $\hat{R}_2$, the numerator consists of (i) the number of interviews obtained at HU's that completed all possible interviews and were also victimized at least once, plus (ii) the number of interviews obtained at HU's for which one of the two interviews was missing and which also reported at least one victimization times a correction factor. This correction factor is used to adjust for the fact that HU's for which an interview is missing may or may not have been victimized during the months covered by the missing interview.

The second estimator, $\hat{R}'_2$ is similar to $\hat{R}_2$ except that the HU's which are missing the first interview and were victimized at least once are multiplied by a different correction factor than the HU's which are missing the second interview and were victimized at least once.

We note that the Bureau of Justice Statistics actually calculated these rates using the weighted counts referred to in Section 4 while we have calculated them using the unweighted counts. In addition, we have used data from unbounded interviews and series as well as individual victimizations. For comparative purposes we compute $\hat{\theta}_2 = 1-\hat{R}_2$ and $\hat{\theta}'_2 = 1-\hat{R}'_2$ to estimate the proportion of crime-free HU's.

Estimator 3: Homogeneous Bernoulli Model

Estimator 3 is based on a homogeneous Bernoulli model of victimization. Let

$$x_{ij} = \begin{cases} 1 & \text{if HU } i \text{ is victimized at least once in month } j \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1,\ldots, H$; $j = 1,\ldots, 12$, where H is the total number of HU's in the sample, be independent Bernoulli random variables with a common value of $p = \Pr\{x_{ij}=1\}$.

Under this model, every HU has the same probability, p, of being victimized in any month. A given HU is victimized in month j independently of whether or not it is victimized in month $\ell(\ell\neq j)$, and HU i is victimized independently of HU k $(k\neq i)$.

If we assume that the missing data are missing at random, it is easily shown (see Eddy, Fienberg, and Griffin, 1981) that the maximum likelihood estimator of p is

$$\hat{p} = \frac{V}{T}, \qquad (3)$$

where T is the total number of x_{ij}'s that are observed and V is the sum of the observed x_{ij}'s, that is, V is the total number of months observed in which a victimization occurred.

If we wish to estimate θ, the probability that a HU is crime-free for the year, then we need only note that

$$\theta = (1-p)^{12}, \qquad (4)$$

and thus the maximum likelihood estimator of θ is

$$\hat{\theta}_3 = (1-\hat{p})^{12} = \left(\frac{T-V}{T}\right)^{12}$$

$$= \left(\frac{\text{\# of crime-free months observed}}{T}\right)^{12} \qquad (5)$$

Estimator 4: A Correlated Bernoulli Model

Tallis (1962) discusses a model which is a "mixture" or weighted average of the model of independence of victimizations across months described above and the model of perfect correlation in which a HU that is victimized in January is victimized every month and a HU that is not victimized in January is never victimized. In this model p represents, as above, the probability that an HU is victimized at least once in a given month. A second parameter, ρ $(0\leq\rho\leq1)$, represents the correlation between any two months of data for a given HU. In particular, letting x_{ij} be defined as above, we suppose that $(x_{i1},\ldots, x_{i12})$, $i = 1,\ldots, H$, follow the Tallis model and that $(x_{i1},\ldots, x_{i12})$ is independent of $(x_{k1},\ldots, x_{k12})$ for $i\neq k$.

As with the homogeneous Bernoulli model, every HU has the same probability p of being victimized in any month, and HU i is victimized independently of HU $k(k\neq i)$. This model has the feature that, when $0<\rho\leq1$, x_{ij} is no longer independent of $x_{i\ell}$ but when $\rho=0$ this model simplifies to the Bernoulli model with complete independence of monthly observations.

Assuming that the missing data are missing at random, we can calculate the likelihood function in terms of p and ρ (details are given in Eddy, Fienberg, and Griffin, 1981). Unfortunately, this function cannot be maximized directly and iterative methods are required to obtain maximum likelihood estimates of p and ρ. Two views of this likelihood are shown for an example in Appendix II. Once we compute these estimates, it is then straightforward to compute the maximum likelihood estimate of θ, the proportion of crime-free HU's:

$$\hat{\theta}_4 = (1-\hat{\rho})(1-\hat{p})^{12} + \hat{\rho}(1-\hat{p}) \quad . \qquad (6)$$

Note that $\hat{\theta}_4$ is a linear combination of $\hat{\theta}_3$, the Bernoulli estimator, and $(1-\hat{p})$, the estimator that results from perfect correlation.

Estimator 5: A Markov Model

Let $x_{i1}, \ldots, x_{i12}$ be defined as above. We can consider these observations as arising from a two state Markov chain with states 0 and 1, where 0 indicates that no victimization occurred within the month and 1 indicates that at least one victimization occurred. Let (p_0, p_1) be the initial probability vector, e.g. p_0 is the probability of no victimization in the initial month, and

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}. \qquad (7)$$

the transition matrix, e.g. p_{00} is the probability of no victimization in month i+1 *given* no victimization in month i. In addition, we assume that HU j is victimized independently of HU k and that the missing data are missing at random. As in the previous model, we are unable to maximize the corresponding likelihood explicitly, but iterative methods can be used to obtain maximum likelihood estimates of p_0, p_{00}, and p_{10}, i.e. $\hat{p}_0$, $\hat{p}_{00}$, and $\hat{p}_{10}$. Several plots of the likelihood function are shown for an example in Appendix II. The maximum likelihood estimate of θ, the proportion of crime-free housing units, is then

$$\hat{\theta}_5 = \hat{p}_0 \, (\hat{p}_{00})^{11} \, . \qquad (8)$$

Note that $\hat{\theta}_5$ represents the probability of starting in the crime-free state in January and moving from the crime-free state to the crime-free state for each of the eleven successive months.

6. Empirical Results: Preliminary Estimates for Victimization Prevalence

Using the sample of 1539 household locations described in Section 3, we have calculated values for the seven estimators presented in Section 5. In addition, by assuming that every nonresponse month was a month in which a victimization occurred, we can compute a lower bound for $\hat{\theta}_1$. Similarly by assuming that every non-response month was a crime-free month, we can calculate an upper bound for $\hat{\theta}_1$. While these bounds are strictly applicable to only $\hat{\theta}_1$, they are quite informative, and suggest the range of possible estimates of θ that can result from changes in model assumptions and specifications. In Table III, below, we display the seven estimates and the bounds for the three years for which complete longitudinal data are available, 1973 - 1975. The estimated standard errors for several of these estimates are available, but not reported here.

TABLE III

Estimated Proportions of Households Untouched
by Crime, 1973 - 75

	1973	1974	1975
Upper Bound	.753	.760	.770
$\hat{\theta}_1$	.706	.719	.732
$\hat{\theta}'_1$	.676	.672	.695
$\hat{\theta}_2$	.672	.689	.689
$\hat{\theta}'_2$	.671	.687	.687
$\hat{\theta}_3$	.580	.583	.626
$\hat{\theta}_4$	.626	.619	.651
$\hat{\theta}_5$	.606	.613	.660
Lower Bound	.497	.519	.523

We note that $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}'_2$, $\hat{\theta}_3$ and $\hat{\theta}_5$ show a small increase from 1973 to 1974 and increase or remain the same from 1974 to 1975. On the other hand, $\hat{\theta}'_1$ and $\hat{\theta}_4$ decrease slightly from 1973 to 1974 and then increase from 1974 to 1975. All the changes are of relatively small magnitude. That is, the proportion of crime-free HU's seems to remain fairly constant over the 1973 to 1975 time period.

In addition, we note that the "upper bounds" are relatively close to the values of $\hat{\theta}_1$, which we believe to be an overestimate of θ. The "lower bounds" lie far from $\hat{\theta}_3$, primarily because they treat a situation with a low value of p (the probability of victimization in given month) as having occurred everytime we have missing data (something that happens quite often). Nonetheless what we can learn from Table III is a rough range for θ (somewhere between 0.5

and 0.8).

7. Extensions and Problems for Further Study

This paper has described some initial attempts to develop models for the analysis of longitudinal files constructed from a rotating sample survey. Our focus has been on the implications of such modelling for aggregate cross-section-like quantities -- in this instance annual victimization prevalence rates for household locations (HU's). We do not believe that modelling NCS data longitudinally at the HU level makes very much sense. Thus the empirical results we report in Section 6 are intended for illustrative purposes only. Even so, the estimates of the prevalence-related parameter θ reported there are clearly overestimates. This is because an HU can report information for a given 6-month period in the NCS, but individuals within that HU can be nonrespondents and we have no information about their possible victimization experiences.

What should be clear from the discussion in the present paper is that the missing data present a far greater problem for rotating panel surveys than has been acknowledged by those who conduct them. Indeed, reports of "monthly non-response rates" of 4-5% for the NCS give no clue to the magnitude of the missing problem that awaits the survey analyst who approaches survey data files organized longitudinally.

In Section 5 we modelled the missing data as if they were missing at random. In fact, we believe that missingness may well be related to victimization experiences, and thus attention needs to be given to modelling missingness and victimization simultaneously.

Finally, we recall that we were able to act as if we had a simple random sample of HU's, even though such an assumption was inappropriate for the full longitudinal data file. Future modelling efforts will need to consider how the sample design characteristics should be reflected in the modelling and analysis process.

8. Acknowledgments

This material is based in part on work performed under Contract No. J-LEAA-015-79 with the Bureau of Justice Statistics, and in part on subsequent work performed under Grant 81-IJ-CX-0087 from the National Institute of Justice, both in the Office of Justice Assistance Research and Statistics, U.S. Department of Justice. Points of view and opinions stated herein are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Bibliography

Alexander, C.H. (1981). "Plans to produce National Crime Survey estimates of 'residences touched by crime'". Unpublished Census Bureau memorandum.

Eddy, W.F., Fienberg, S.E., and Griffin, D.L. (1981). "Estimating victimization prevalence in a rotating panel survey." Bulletin of the International Statistical Institute. Forthcoming.

Fienberg, S.E. (1978). "Victimization and the National Crime Survey: Problems of design and analysis." In *Survey Sampling and Measurement* (N.K. Namboodiri, ed.), New York: Academic Press, pp. 89–106.

Fienberg, S.E. (1980a). "Statistical modelling in the analysis of repeat victimization." In *Indicators of Crime and Criminal Justice: Quantitative Studies* (S.E. Fienberg and A. Reiss, Jr., eds.), Washington, D.C.: U.S. Government Printing Office, 54–58.

Fienberg, S.E. (1980b). "The measurement of crime victimization: prospects for panel analysis of a panel survey." *The Statistician* 29, 313–350.

Kalbfleisch, J.D. and Prentice, K.L. (1980). *The Statistical Analysis of Failure Time Data.* New York, John Wiley and Sons.

Little, R. (1980). "Superpopulation models for non-response II: The non-ignorable case." National Academy of Sciences, Committee on National Statistics Panel on Incomplete Data. To be published.

Penick, B.K. and Owens, M.E.B. (eds.) (1976). *Surveying Crime* (Report of Panel for the Evaluation of Crime Surveys), Washington, D.C.: National Academy of Science.

Reiss, A.J. Jr. (1980). "Victim proneness by type of crime in repeat victimization." In *Indicators of Crime and Criminal Justice: Quantitative Studies* (S.E. Fienberg and A. Reiss, Jr., eds.), Washington, D.C.: U.S. Government Printing Office, 41–53.

Rubin, D.B. (1976). "Inference and missing data." *Biometrika* 63, pp. 581–592.

Tallis, G.M. (1962). "The use of a generalized multinomial distribution in the estimation of correlation in discrete data." *J. Roy. Statist. Soc. B.* 24, 530–534.

U.S. Department of Commerce, Bureau of the Census (undated). *National Crime Survey, National Sample, Survey Documentation.*

U.S. Department of Justice, Bureau of Justice Statistics (1980). *Criminal Victimization in the United States, 1978.* Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice, Bureau of Justice Statistics (1981a). "The prevalence of crime." *Bureau of Justice Statistics Bulletin.* Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice, Bureau of Justice Statistics (1981b). *Criminal Victimization in the United States, 1979.* Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice, Bureau of Justice Statistics (1981c). *National Crime Surveys: National Sample, 1973 - 1979.* Ann Arbor, Michigan: Inter-University Consortium for Political and Social Research, University of Michigan.

Appendix I: The BJS Estimators

In order to describe the BJS estimators, Alexander (1981) presents a classification of each HU into one of six groups according to the number of interviews and the types of noninterview. There are three types of noninterview. A type A noninterview occurs when household members are rarely at home, uncooperative, or otherwise impossible to reach. A type B noninterview occurs when an HU selected for sample turns out to be vacant or otherwise ineligible. A type C noninterview occurs when an HU is found to be demolished, converted to non-residential use or otherwise out of the scope of the NCS. The six groups are as follows:

group a: both records are interviews.

group b: only the first record is an interview -- the second record is missing because the HU was rotated out of the sample, or the second interview is a type A noninterview.

group c: only the second record is an interview -- the first record is missing because the HU has just been rotated into the sample, or the first record is a type A noninterview.

group d: the first record is an interview; the second record is a type B or C noninterview.

group e: the first record is a type B or C noninterview; the second is an interview.

group f: neither record is an interview.

From these groups, the following quantities are computed:

H1 = # of HU's in group a.

H4 = 1/2(# of HU's in group b).

H5 = 1/2(# of HU's in group c).

H6 = 1/2(# of HU's in group d).

H7 = 1/2(# of HU's in group e).

C1 = # of HU's in group a that report at least one victimization in either interview.

C2 = 1/2(# of HU's in group a that report at least one victimization in the first interview).

C3 = 1/2(# of HU's in group a that report at least one victimization in the second interview).

C4 = same as C2 but for group b.

C5 = same as C3 but for group c.

C6 = same as C4 but for group d.

C7 = same as C5 but for group e.

The BJS victimization prevalence rates are then given by:

$$\hat{R}_2 = \frac{C1 + C6 + C7 + (C4+C5)\ [C1/(C2+C3)]}{H1 + H4 + H5 + H6 + H7}$$

and

$$\hat{R}'_2 = \frac{C1 + C6 + C7 + C4[C1/2C2] + C5[C1/2C3]}{H1 + H4 + H5 + H6 + H7}$$

Appendix II: Likelihood Plots

In order to obtain a more complete understanding of the maximum likelihood estimates, presented in Section 5, for the correlated Bernoulli model and the Markov model, we have generated several graphs of the likelihood surface of each of these two models.

Figures 1 and 2 show two perspectives of the loglikelihood surface for the correlated Bernoulli model for the 1973 data. We can see that the surface has a unique maximum and so the maximum likelihood estimates are unique. Near the maximum the surface is much more peaked with respect to the variable p than with respect to ρ. In fact, from Figure 2, we note that, near the maximum likelihood estimate of p, the surface is very nearly flat with respect to ρ. Thus the variance of $\hat{\rho}$ is large compared to the variance of $\hat{p}$. The loglikelihood surfaces for 1974 and 1975 data are similar to the one for 1973, as illustrated by Figures 1 and 2.

Figures 3 to 7 are views of the likelihood surface for the Markov model with one parameter set equal to its maximum likelihood estimate. These 5 figures were generated using the 1975 data.

In Figures 3 and 4 p_0, the probability of initially being in the state 0 (non-victimized), is set to the value of $\hat{p}$ = .954.. This loglikelihood surface is curved with respect to p_{00} much more than with respect to p_{10} and so the variance of $\hat{p}_{00}$ will be small relative to the variance of $\hat{p}_{10}$. This is as we would have expected since we have many more observations in the state 0 than in the state 1, and hence have more information about 0 to 0 transitions than about 1 to 0 transitions. Although it is not readily seen from these two figures, the maximum is unique.

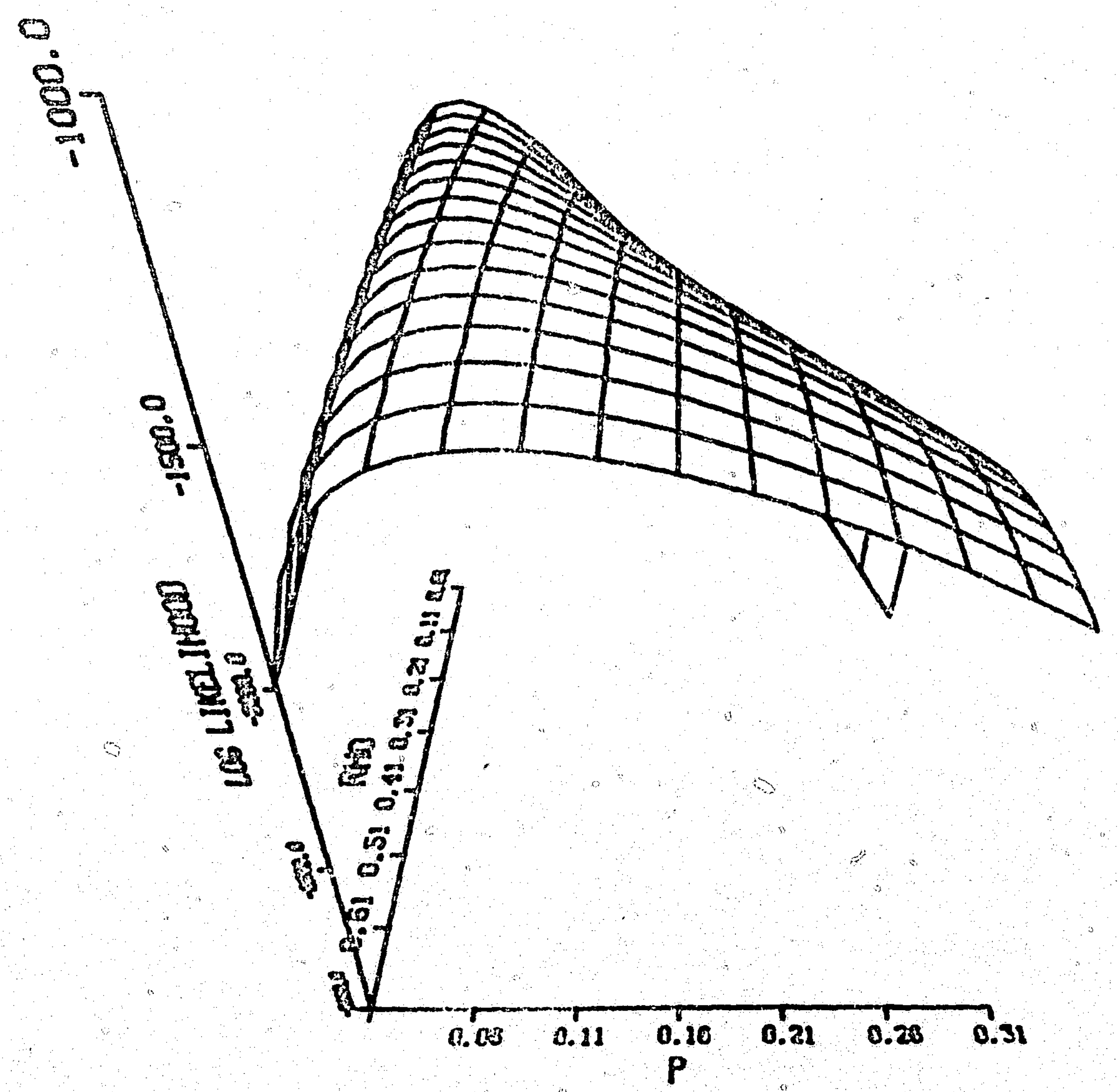
Figures 5 and 6 show the loglikelihood surface with p_{10}, the probability of moving from state 1 to state 0, fixed at its maximum likelihood value of .838. We see that the surface is more curved in the p_{00} direction than in the p_0 direction. This is, again, what we would have expected since there are many more 0 to 0 transitions than initial observations of the state 0. In these 2 figures it is possible by close inspection to see that the maximum is unique.

In Figure 7 p_{00}, the probability of moving from state 0 to state 0, is fixed to be .967. We see that this loglikelihood surface is extremely flat. Thus the va-iances of $\hat{p}_0$ and $\hat{p}_{10}$ will be relatively large and the values of $\hat{p}_0$ and $\hat{p}_{10}$ may not be very informative. It is vaguely discernable that a unique maximum occurs in the upper right hand portion of the figure.

The loglikelihood surfaces for the Markov model using data from 1973 and 1974 have characteristics similar to those displayed in Figures 3 to 7.

LOG LIKELIHOOD FOR 1973 DATA WITH SERIES

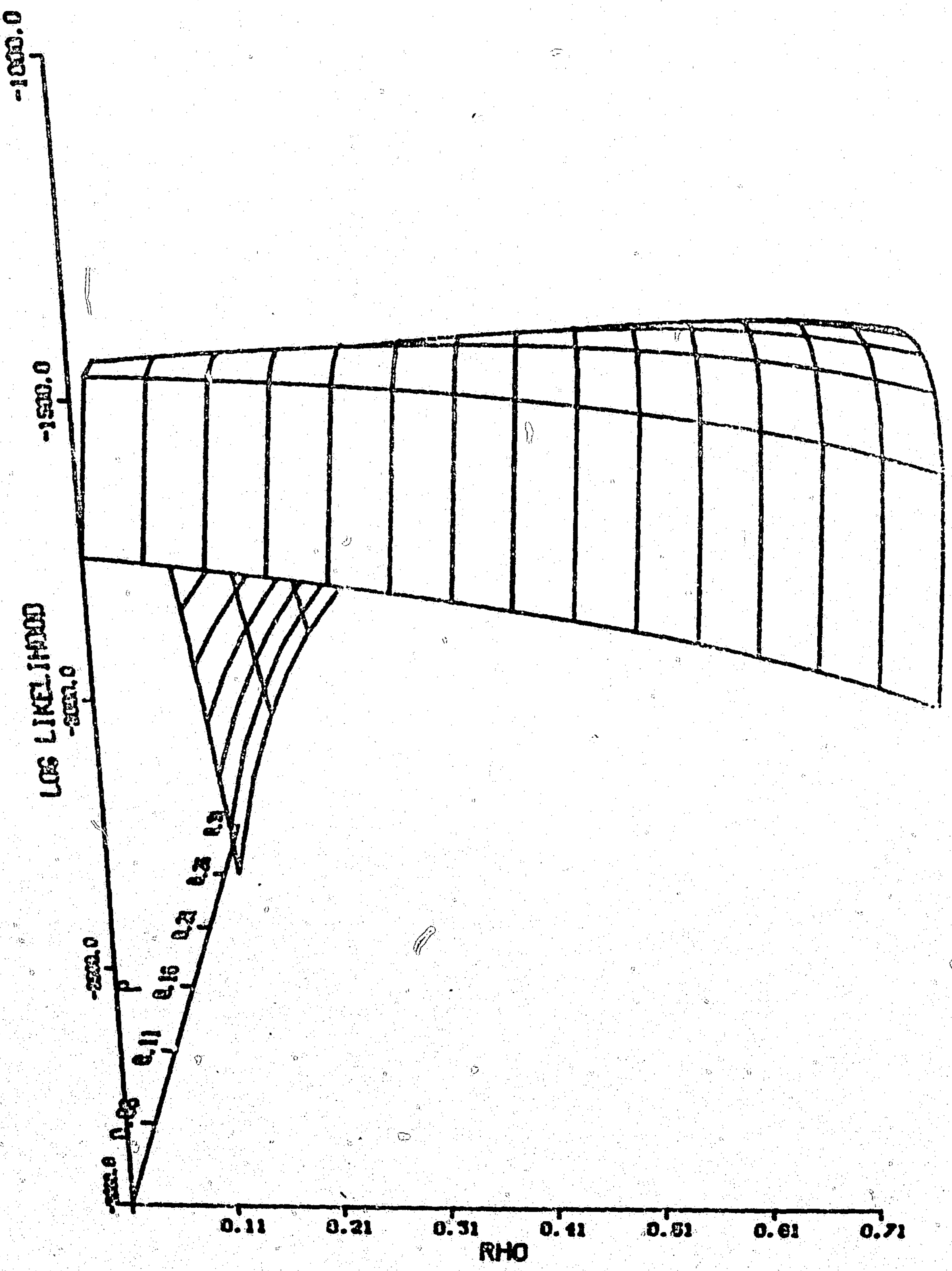
Figure 1



Preceding page blank

LOG LIKELIHOOD FOR 1973 DATA WITH SERIES

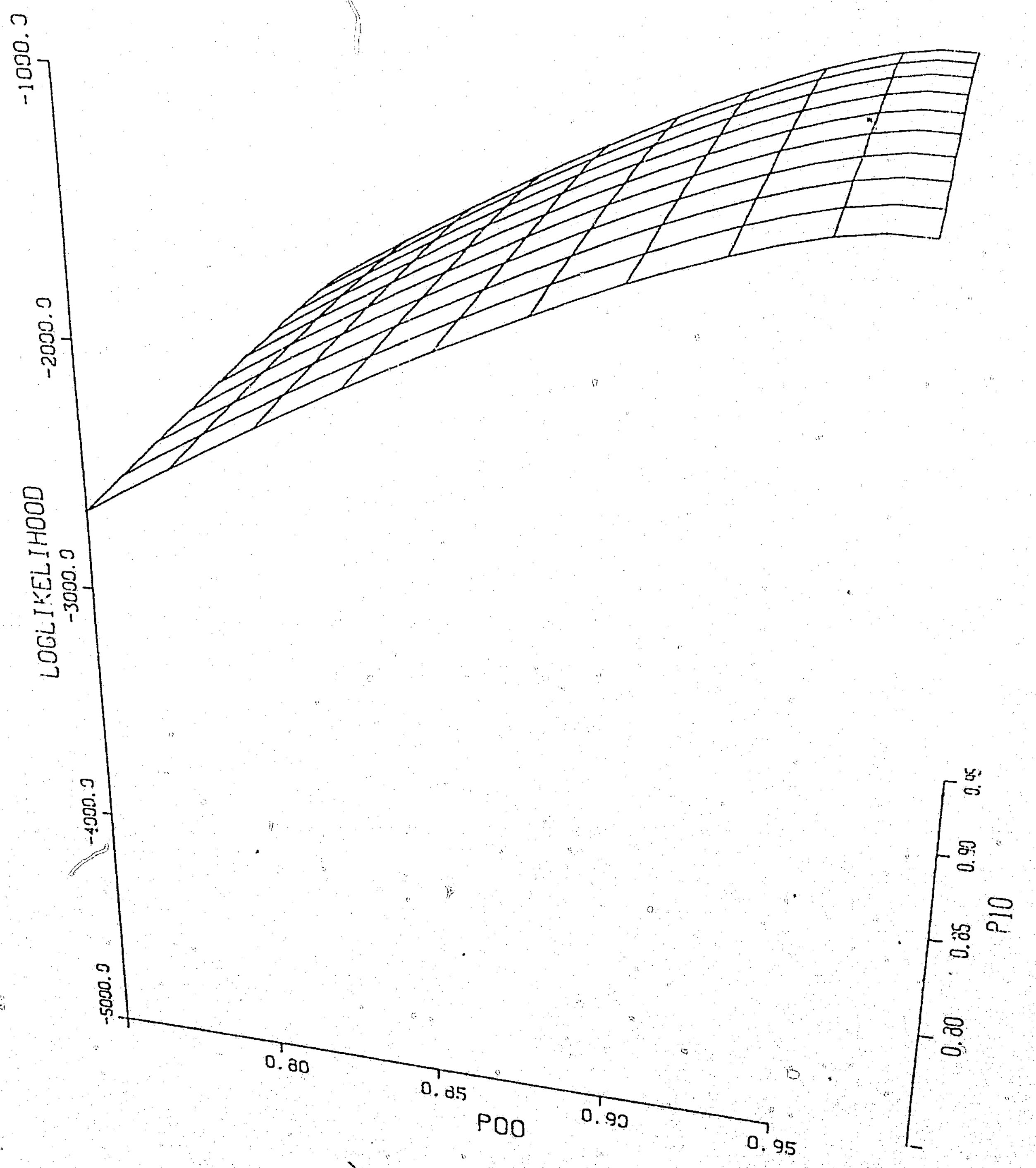
Figure 2



LOG LIKELIHOOD FOR MARKOV MODEL

Figure 3

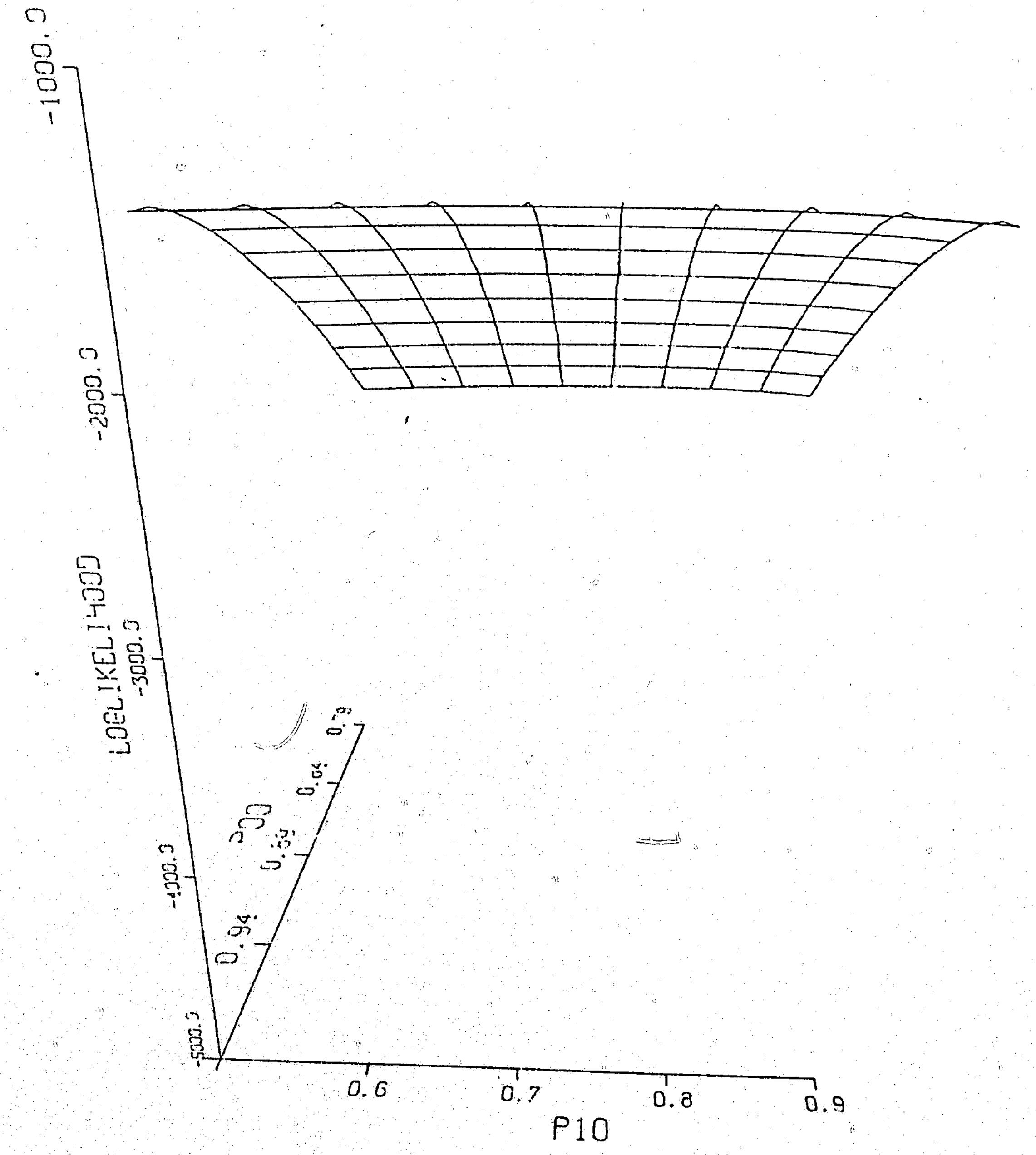
1975S PO=.954



LOG LIKELIHOOD FOR MARKOV MODEL

Figure 4

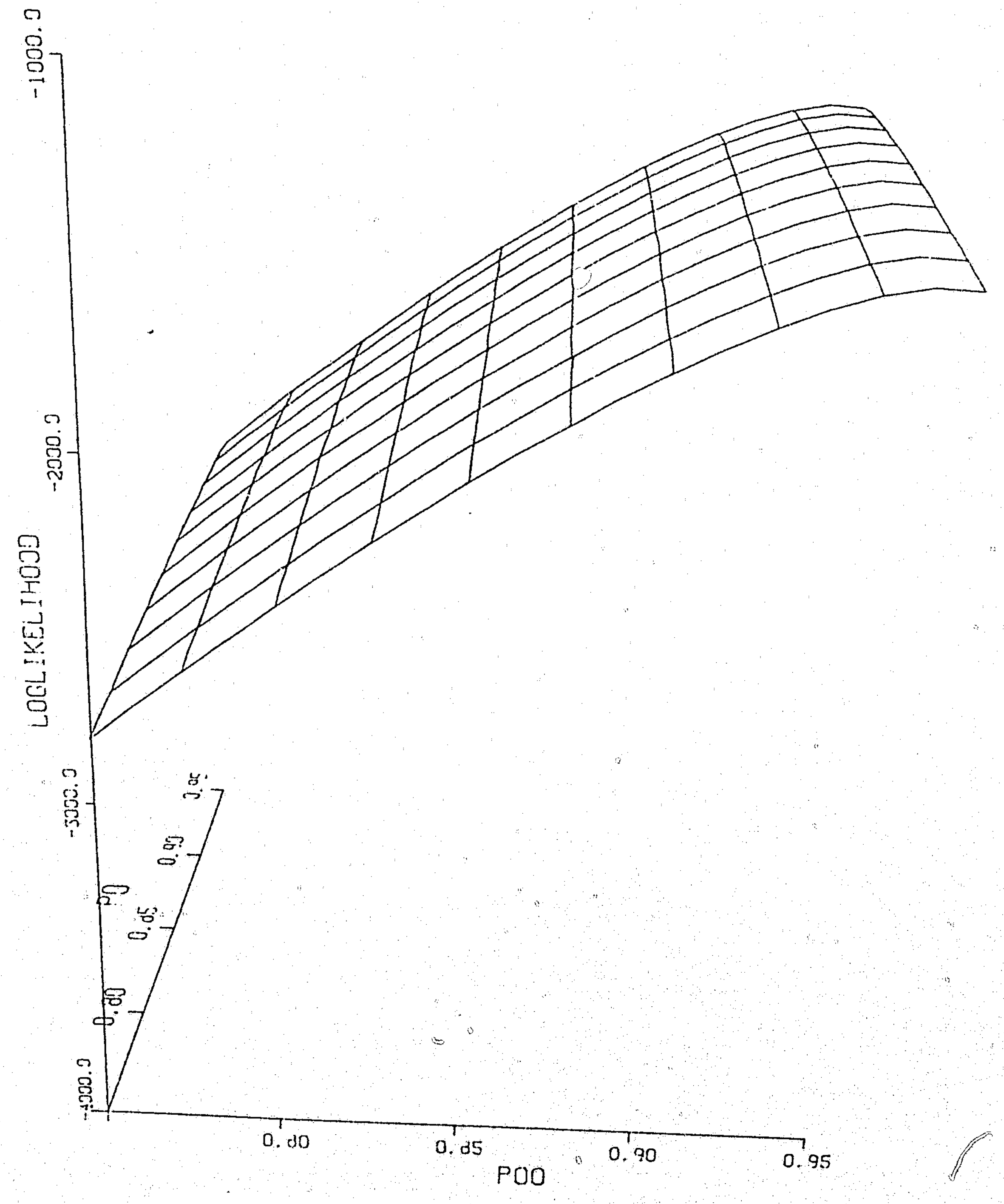
1975S PO=.954



LOG LIKELIHOOD FOR MARKOV MODEL

Figure 5

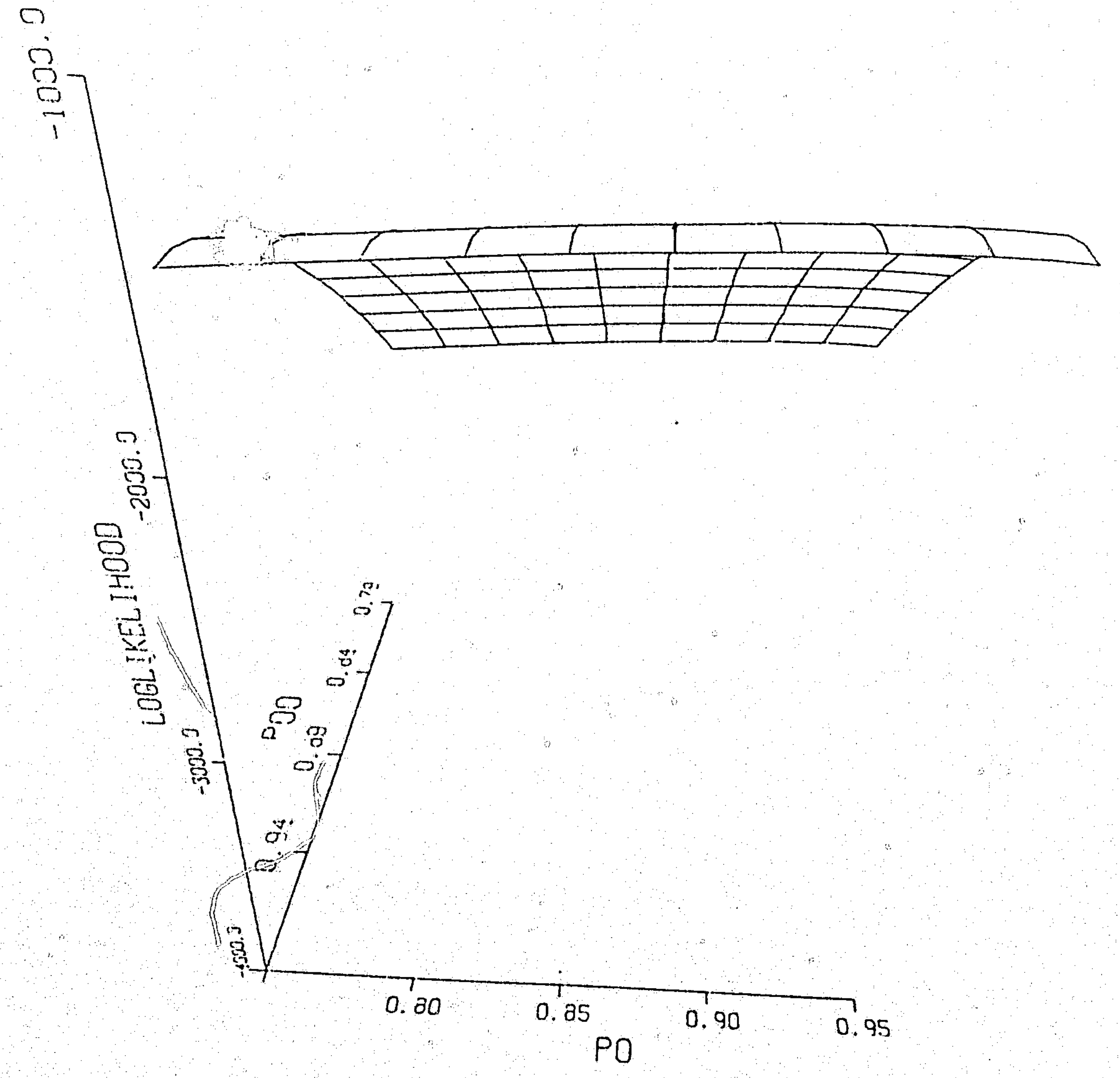
1975S P10=.838



LOG LIKELIHOOD FOR MARKOV MODEL

Figure 6

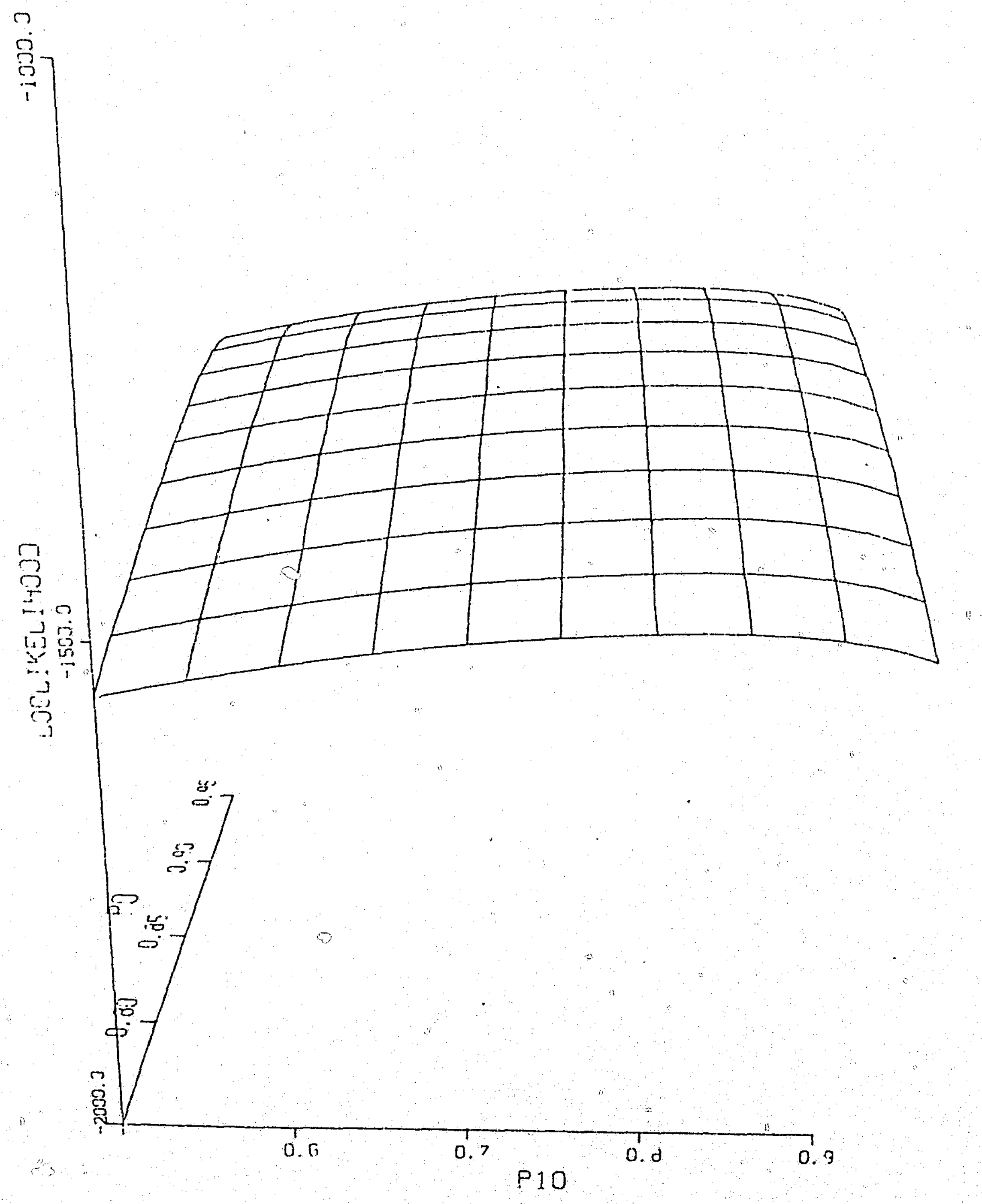
1975S P10=.838



LOG LIKELIHOOD FOR MARKOV MODEL

Figure 7

1975S   P00=.967



END