General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

GRADUATE INSTITUTE OF STATISTICS

TEXAS A&M UNIVERSITY · COLLEGE STATION

CR-115201

FINAL TECHNICAL REPORT

June 1, 1969 to May 31, 1971

COMPARTMENTAL ANALYSIS AND MISSING DATA TECHNIQUES IN STATIONARY TIME SERIES

National Aeronautics and Space Administration

Manned Spacecraft Center

Grant #NGR 44-001-097

W. B. Smith
Associate Professor
Institute of Statistics
Texas A&M University
College Station, Texas 77843

FINAL TECHNICAL REPORT

COMPARTMENTAL ANALYSIS AND MISSING DATA TECHNIQUES JN STATIONARY TIME SERIES

1. Introduction.

Hocking and Smith [1968] have presented a technique for estimating the parameters of a multivariate normal distribution from data vectors, some of which have missing elements. A purpose of this grant was to extend this technique to estimation problems of biological interest. The procedure uses all available data, both full and partial vectors, and can be outlined as follows: (1) group the data according to which variates are missing, (2) estimate within each group the available parameters, (3) combine in a sequential fashion (optimally) the estimates of part (2).

The procedure has been shown to possess some optimality problems and, in fact, the combination of parameter estimates was done in such a way as to minimize its variance and to retain unbiased properties if possible.

Section 2.1 outlines in some detail the procedure and its extensions, both theoretical and to biological problems, carried out under this grant. It should be noted that in addition to the biological problems of interest other applications are immediately evident. For an example missing and/or partial data records can occur due to random machine or telemetry failure from a spacecraft to the ground or on the ground. Thus the procedures developed under this grant can be used to salvage those partial data records. That is, rather than discarding partial data records the information available within them is utilized without resorting to estimating the missing data elements.

Another purpose of this grant was to develop and extend techniques for biological applications of compartmental analysis. Many biological and physical systems can be modelled mathematically by a series of compartments. As examples, the digestive system of mammals or the flow of interaction of documents in a research lab can be thought to be sequences of compartments. The mathematical analysis of such models have been limited to a deterministic situation, that is, one in which the particles within the compartments flow at specific rates. Research on this grant considers the sequence of compartments and the flow between compartments in a stochastic (random) fashion. That is, within each compartment a particle will possess a probability distribution which relates the chances of its exiting the compartment at a particular time. Thus, the flow rates (time dependent) between compartments are of interest. As an example one could think of a mathematical model of an epidemic. That is, the organism would be within a single individual, each individual having a susceptible period and a period after which he contracts the disease. The time interval in which the organism resides within the individual can itself be broken down into a latent period (a period in which no symptoms are observable), an infective period (a period in which it may or may not exhibit symptoms but in which the individual can pass the organism to another individual), and a recovery period. Each of these periods obviously vary in some fashion from individual to individual depending upon the health of the individual, the severity of the case involved, and other extraneous conditions. Thus it would be obvious that a deterministic model would not fit such a situation whereas a stochastic model could. We describe in Section 2.3 solutions found to specific types of stochastic compartmental models both for the estimation and the model itself.

In Section 2, brief descriptions of each technical advance achieved under this grant will be given. The research carried out under the auspices of this grant has received good reviews by colleagues, as is evidenced by their publication in the respected journals. The lag in publication time prevents us from reporting the final dispensation of two of the technical reports. In total there were six technical reports issued under the grant, four of which have been published and the other two submitted for publication.

2. Review of Results.

2.1. Missing Data.

Technical Report #1 entitled "Selection Index Estimation from Partial Multivariate Normal Data", represents an extension of the Hocking-Smith technique to the estimation of the well known selection index. The index is based upon a multivariate time dependent data vector and constitutes a legitimate extension of the partial data technique developed by Hocking and Smith. The procedure is spelled out in considerable detail in the technical report but can be summarized:

(1) The procedure is applied to the estimation of the selection index (linear) when incomplete multivariate normal data vectors are available. (2) The procedure utilizes all data available (both full and partial data vectors) and represents an improvement in the precision of the estimator over that found by using the full data vectors only. (3) Estimates of the phenotypic means and covariance matrix are also found as a byproduct. (4) Individuals with partial data vectors are indexed by an extension of the technique and this index is compared to that proposed by Henderson [1962]. (5) The results of Monte Carlo computer simulations are also tabulated.

That is, defining $I_j = b'X_j$, where I_j is a composite index value associated with the j^{th} member of the population, b is an n x 1 vector of unknown coefficients, and X_j is n x 1 vector of phenotypic observations on the j^{th} member of the population, Smith [1936] shows that $b = P^{-1}G\alpha$. In this context P is the n x n variance-covariance matrix of phenotypic values, G is the n x n variance-covariance matrox of genotypic values, and α is an n x 1 vector of weights associated with the n traits. The problem of estimating b is attacked by estimating P, the matrix of phenotypic values. This extension to both full and incomplete data vectors represents the first use of an index on partial data.

Further extensions are given in Technical Report #5 in which the parent distribution of the phenotypic observations vector is multinomial. Again both full and partial data vectors are used, all information being gleaned from them. The phenotypic mean vector μ and the phenotypic variance-covariance matrix P are also estimated by the process. Individuals with partial data vectors are indexed. For further details of both the multivariate normal selection index and the multinomial selection index see Technical Reports #1 and #5, respectively.

The selection index has been quite useful in selecting on animals in a population. Further applications are to the selection of pilots for test purposes by Aerospace Research Labs at Wright-Patterson AFB, Dayton, Ohio. In addition genetic applications are obvious.

2.2. Time Series.

A time series application is contained in Technical Report #4. Consider the classical minimum variance unbiased estimator of a population mean in a discrete parameter, covariance stationary, stochastic process when the data is sampled by rotation sampling, as derived by Yates [1949] and Patterson [1950]. This estimator is invariant under alterations in the rotation scheme. In Technical Report #4 we derive by a constrained optimization procedure similar simultaneous equations. Patterson has considered the technique but rejected it because of the lengthy estimator expressions, however by introducing matrix expressions and assuming a special rotation scheme this technique is now to be tractable. That is, we give an alternate way of deriving the estimator.

The merits of the procedure are as follows: (1) Exact expressions from a population mean estimator can be found on each occasion. (2) The matrix notation leads to easier computer programming. (3) The derivation of the exact expression of the variance-covariance matrix leads to ease in investigating the large sample properties of the estimator and their relation to maximum likelihood.

The Yates-Patterson minimum variance unbiased linear estimate of the population mean on the last observation occasion was made under the following assumption: (1) The correlation coefficient between observations in the same unit i occasions apart is ρ^i and is known, $1 \leq i \leq h$. (2) The variances are the same on each occasion and are known (they are noted by σ^2). (3) Sample units on each occasion are equal to n. (4) $n\phi$ units, $0 \leq \phi \leq 1$, are replaced by newly chosen units on each occasion and (5) sampling on each unit is done mutually independently from an infinite population, so that the correlation coefficient between two observations on different units is zero. The resulting estimator is given by

$$Y_h = \varphi_h \bar{y}''_h + (1 - \varphi_h) \{\bar{y}'_h + \rho(Y_{h-1} - \bar{x}'_{h-1})\} \tag{1}$$

where $\bar{x}'_{h-1}$ is the mean of observations on occasion h-1 associated with the $n(1 - \phi)$ units common to occasion h, $\bar{y}'_h$ is the sample mean on occasion h associated with the same common units, $\bar{y}''_h$ is the mean on occasion h associated with the newly chosen uncommon units, and Y_{h-1} is the estimator of $\mu(h-1)$ (the mean at occasion h-1) based on the observation up to occasion h-1. Now φ_h satisfies

$$\varphi_h = \frac{\rho^2(1 - \phi)\varphi_{h-1} + (1 - \rho^2)\phi}{\rho^2(1 - \phi)\varphi_{h-1} + (1 - \rho^2)\phi + (1 - \phi)} \tag{2}$$

and

$$\text{var } (Y_h) = \varphi_h \sigma^2/n\phi \quad \text{given that } \varphi_1 = 1 . \tag{3}$$

Patterson further derived the minimum variance unbiased linear estimate of $\mu(h-k)$ when all observations up to the h occasion are available. We derived a minimum variance unbiased linear estimate of each of the above population means by constrained minimization of the variance of the linear estimate with respect to its coefficient. That is, we derive these estimates by minimizing the variance while guaranteeing unbiasedness. The result of this procedure uses Aitken's generalized least square estimator. A simplified specification of the estimator is given and the variance-covariance matrix is spelled out. For further details into the technical properties of the procedure, as well as, a specific formulation of the estimator itself, see Technical Report #4. A revised version of Technical Report #4 has appeared in the Report on Statistical Applications published by the Japanese Union of Scientists and Engineers, March, 1971.

2.3. Compartmental Analysis.

The introduction of compartmental analysis has stirred considerable interest among bio-mathematicians in modelling biological clearance, that is, the passage of a given material through a biological system. Virtually all current investigations are designed such that the system requires a tracer material to enter either by a large initial pulsing or by small continuous dosing. The initial pulse method is typically easier to administer physically, while the continuous dosing has an advantage of containing some checks on the underlying compartmental assumption (Hearon, [1968]). Both have been used to estimate turnover rates in the deterministic system.

Technical Report #2 "Stochastic Compartmental Analysis: Model and Least Squares Estimation from Time Series Data", by J. H. Matis and H. O. Hartley introduced a stochastic behavior to compartmental systems which are pulse labelled. A homogeneous partial differential equation which characterized the stochastic behavior is solved for distribution theory, and the distribution theory in turn provides a basis for the recommended estimation procedure.

Other modelling of biological clearance has been based primarily on the idea of compartmental analysis assuming linear kinetics; the data has been fitted to sums of negative exponentially distributed random variables with subsequent interpretations relative to the number of compartments turnover rates, etc. Recently, Wise, et.al. [1968] have derived an alternate model based on powers of time, and they have exhibited a great many fits of the data to their model. Our research contributes a generalization to the standard compartmental model by incorporating age dependency into the system. The generalization not only includes the standard compartmental model and the power of times model, in special cases, but also strengthens the power of time model by interpreting it within the compartmental framework. Following the lines of Technical Report #2, Technical Reports #3 and #6 present research introducing a gamma sojourn time distribution in the place of the previous negative exponential distribution. The resulting data dependency of turnover rates then induce the powers of time. A biological example included in this research demonstrates the analytical recognition of age dependency phenomena and the estimation procedure of the age dependency model.

For further details on the specific mathematical formulations of the procedures outlined above see Technical Reports #2, #3 and #6 issued under this grant and the papers published as results of modifications of the above technical reports.

3. Summary.

In summary this grant has supported important research in biological statistics. Each of the three main areas outlined in Section 2 yielded some interesting and important results. Immediate applications of each of the resulting techniques to biological and physical situations have resulted. Widespread interest in the results is evidenced by the significant number of reprint requests received by the authors.

Selection index estimation utilizing the distributional properties of the input variates is of interest to geneticists, in particular, and other scientists involved in many varieties of selection problems. It was with some marked interest that the principal investigator discovered that research on the selection index and its application to pilot selection was being carried on at the Aerospace Research Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio by Dr. David A. Harville.

The compartmental analysis applications to many problems are currently being carried out. One such problem only being investigated at its very beginning stages is an application of stochastic compartmental analysis through the estimation of the flow rate of certain biological micro-organisms, which would yield information concerning the degree of contagious infection resulting from each. During deep space probes the extended confinement by a small group of space travellers would bring each in contact with the other disease potentials. Thus if such an estimation procedure were available before such a deep probe one would be able to simulate on a computer the possibilities of certain contagious diseases occurring within the crew. Further research on this problem will hopefully be carried out under alternate funding.

4. Technical Reports.

Report #1, "Selection Index Estimation from Partial Multivariate Normal Data", by W. B. Smith and R. C. Pfaffenberger (appeared in Biometrics, 26, 625-640).

Report #2, "Stochastic Compartmental Analysis: Model and Least Squares Estimation from Time Series Data", by J. H. Matis and H. O. Hartley (appeared in Biometrics, 27, 77-102).

Report #3, "Stochastic Compartmental Analysis: Some Applications and Examples in Pulse Labelled Systems", by J. H. Matis and H. O. Hartley and W. B. Ellis (appeared in part in Biometrics with Report #2).

Report #4, "Estimation of the Mean in a Discrete Parameter, Covariance Stationary, Stochastic Process in Rotation Sampling", by Kin-Ya Nishikawa and W. B. Smith (appeared in Report on Statistics Applications Research, Japanese Union of Scientists and Engineers, 18, No. 1, 9-21).

Report #5, "Multinomial Selection Index", by W. B. Smith and D. M. Scott, (submitted for publication to Biometrics).

Report #6, "An Example of Age Dependency in Compartmental Analysis", by J. H. Matis (submitted for publication to Biometrics).

5. References.

Hearon, J. Z. [1968], "A Washout Curve in Tracer Kinetics", Mathematical Biosciences, 3, 31-40.

Henderson, C. R. [1963], "Selection Index and Expected Genetic Advance", NAS-NRC Publication 982, 141-163.

Hocking, R. R. and W. B. Smith [1968], "Parameter Estimation in the Multivariate Normal Distribution with Missing Observations", Journal of the American Statistical Association, 63, 159-173.

Matis, J. H. and H. O. Hartley, "Stochastic Compartmental Analysis: Model and Least Squares Estimation from Time Series Data", Biometrics, 27, 77-102.

Nishikawa, K., and W. B. Smith, "Estimation of the Mean of a Discrete Parameter, Covariance Stationary, Stochastic Process in Rotation Sampling", Report of Statistics Applications Research, Japanese Union of Scientists and Engineers, 18, No. 1, 9-21.

Patterson, H. D. [1950], "Sampling on Successive Occasions with Partial Replacement of Units", Journal of Royal Statistical Society, Series B, 12, 241-255.

Smith, H. F. [1936], "A Discriminant Function for Plant Selection", Ann. Eugen. London, 7, 240-250.

Smith, W. B. and R. R. Hocking [1968], "A Simple Method for Obtaining the Information Matrix for a Multivariate Normal Distribution", The American Statistician, 22, No. 1, 18-20.

Smith, W. B. and R. C. Pfaffenberger [1970], "Selection Index Estimation from Partial Multivariate Normal Data", Biometrics, 26, 625-640.

Wise, M. E., S. B. Osborn, J. Anderson and R. W. S. Tomlinson [1968], "A Stochastic Model for Turnover of Radio Calcium Based on the Observed Power Laws", Mathematical Biosciences, 2, 199-224.

Yates, F. [1949], Sampling Methods for Censuses and Surveys, London, Charles Griffin.