BEYOND REGRESSION:
NEW TOOLS FOR PREDICTION AND ANALYSIS
IN THE BEHAVIORAL SCIENCES

A thesis presented

by

Paul John Werbos

to

The Committee on Applied Mathematics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

statistics

Harvard University

Cambridge, Massachusetts

August, 1974

PREFACE

The initial impetus for this thesis came from a suggestion by Prof. Karl Deutsch, my thesis supervisor, that I look more closely at the prediction of national assimilation and political mobilization, by use of the Deutsch-Solow model; his comments have been of major help to me with all the empirical work on nationalism, and in revising the original structures of Chapters (V) and (VI). The earlier work reported in section (ii) of Chapter (VI) was carried out under his supervision and support, through a research project funded by the Cambridge Project. I have been surprised more than once by the sensitivity and receptiveness of his intuition, in suggesting areas of research which looked unworkable at first but which led in the end to useful and surprising innovations. The opinions expressed in Chapter (V), however, remain my own responsibility, particularly in their more bull-headed aspects.

Professor Mosteller, in the Harvard Statistics Department, has helped by introducing me to current work on robust estimation, by suggesting the use of simulation tests, and by monitoring the general content of the first four chapters. Professors Anderson and Bossert, of the Committee on Applied Mathematics, and

Professor Dempster, of the Harvard Statistics Department, have helped me to find a more orderly way of presenting the mathematical ideas here, help which was sorely needed in the summer of 1973. To the Cambridge Project, and its sponsors at ARPA, must be given thanks, not only for all the computer time used in this research and in its documentation, but also for the opportunity to translate some of these ideas from theory into operational systems without the extensive delays more common in such research; without their support, this research could never have been brought to a state final enough to allow its continuation.

The personal and financial conditions which led to the completion of this thesis were rather complex, and debts are owed to more individuals than should properly be cited here. Certainly I have a strong debt to my parents in this respect; a debt to Richard Ney, whose ideas on the stock market financed a large part of my activity in this period; a moral debt to Prof. Nazli Choucri, of M.I.T., who, in a brief discussion in the spring of 1973, encouraged me to continue this work; a debt to a colleague, Gopal Krishna, for helping me see more clearly that the psychological hurdles I faced at first were not totally unique; a debt to Dr. Karreman, of the Bockus Research Institute, who, in one of those

1960's programs to encourage high school students, connected with the Moore School of Electronics at the University of Pennsylvania, got me started asking questions about the phenomenon of intelligence, questions which led me to the dynamic feedback concept, which was applied only later to formal statistics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYNOPSIS

This thesis provides a broad, coherent exposition of a new mathematical approach to social studies and to related fields.

This work began as an attempt to apply the classical techniques of statistics and econometrics to the Deutsch-Solow model of nationalism, in order to turn this model into a workable tool for predicting the political future. In the course of this effort, it became clear that the usual statistical methods do a poor job in fitting dynamic models to real-world data, if we judge these models by their ability to make good predictions across time. It also became clear that newer and better methods would not be feasible, economically, unless we could invent less expensive algorithms, too. Thus the goals of this thesis are five-fold: (i) to describe new ways of fitting models to data; (ii) to define new algorithms which make these methods feasible; (iii) to introduce evidence for the superiority of these methods, both for real-world and for simulated data; (iv) to discuss the applications of these ideas, in broad terms, to social and even biological sciences; (v) to discuss the new work on nationalism which has led us in these

directions.

Let us begin with the first three goals.

We have studied not one, but two, new approaches to fitting models to data. First, we generalized the work by Box and Jenkins, on "ARMA" processes, on "mixed AutoRegressive Moving-Average processes." Chapter (III) discusses the mathematics of this approach, in detail. It shows how an ordinary, multivariate autoregressive process, observed by way of "noisy" data (i.e. data measured with random measurement errors or conceptual distortions), becomes a "vector ARMA process." It then shows how to apply "dynamic feedback" to estimate the coefficients of such processes, at much lower costs than were possible previously; the resulting computer program is now available to the public through the MIT Cambridge Project Consistent System. In Chapter (V), we discuss why we considered this approach important for quantitative political science.

In studies of simulated data, the ARMA approach generally yielded only half as much error as regression did, in estimating the coefficients of a simple model; it was more efficient in making use of limited data and it led to less systematic bias, both. However, with real-world data, the ARMA approach did little better

than regression in making long-term predictions; the error distribution curves for ARMA are only about 10% smaller than those for regression, and the curves are uniformly close to each other.

After re-examining these empirical results and the theory of maximum likelihood itself, we formulated a new, more radical and more successful approach to the fitting of models. In essence, the idea is to maximize long-term predictive power directly, over the known data, instead of maximizing formal likelihood. Formally, this idea rests upon the apriori expectation that many social processes are governed by relatively deterministic underlying trends, obscured both by measurement noise and by transient deviations of great complexity. The qualitative, political basis of this idea is discussed in Chapter (V). Sections (vii) and (xi) of Chapter (II) discuss the statistical basis.

The "measurement-noise-only" approach strongly outperformed both ARMA and regression, over both real-world and simulated data. It outperformed ARMA most strongly in our most complex simulated processes, which seem most representative of the real world. According to our error distribution graphs, the new method cuts in half the errors in long-term predictions of real-world variables; the biggest reductions occur

with those variables, such as national assimilation, and with those cases, near the middle of the distributions, for which the simple models of the first half of Chapter (VI) can do an adequate job of prediction.

These empirical results for our new approach came from special computer programs, which exploited the simplicity of the models under study. In Chapter (II), we discuss how the algorithm of "dynamic feedback" can be used to estimate more general "measurement-noise-only" models, at minimal cost, especially for models which are very intricate, nonlinear and nonMarkhovian; we also discuss how the algorithm can fit more conventional models, can optimize policy, and can perform "pattern analysis" - a dynamic alternative to factor analysis. "Dynamic feedback" is essentially a technique for calculating derivatives inexpensively, for use with the classic method of steepest descent. In section (iv) of Chapter (III), and in section (iii) of Chapter (VI), we discuss how our experience here with steepest descent has led us to new ways of adjusting the "arbitrary convergence weights" of steepest descent; these methods speeded up the process of convergence by a large factor. The Appendix to Chapter (II) discusses extensions of these

methods, for the general, nonlinear case.

From a practical point of view, the applications of such mathematical ideas in the social sciences remain controversial. The extreme positions of "behaviorism" and "traditionalism" remain popular; divisions still exist between quantitative and verbal studies of social behavior. In Chapter (V), we describe how our mathematical tools might fit in, in the broader context of social studies and political decision-making. From a utilitarian and Bayesian point of view, we suggest a methodological approach intermediate between "behaviorism" and "traditionalism," in which the different frameworks might be integrated more closely with each other. In sketching out the possibilities for such an integrated framework, we also point out that the algorithms of Chapter (II), taken as part of "cybernetics," have a direct value as paradigms, to help us understand the requirements of the complex information-processing problems faced by human societies and by human brains. We also mention possible applications to other fields, including ecology.

Finally, Chapter (VI) presents our empirical and analytic work on nationalism.

In sections (II) and (III), we discuss our success

in making long-term predictions of national assimilation and mobilization, by use of the Deutsch-Solow model. Table VI-8, for example, gives the average errors in predicting the <u>percentage</u> of population assimilated, over time periods on the order of thirty years; these errors are uniformly distributed between 0% and 2%, except for four outliers (20% of all cases) at 2.68%, 3.08%, 3.09% and 6.21%. The failures of these predictions are also informative; they give us a picture of those external factors which really do have the power to divert the processes of assimilation and mobilization from a steady course. We have tabulated the predictions of the "robust" method for the years 1980, 1990 and 2000; these predictions are subject to caveats discussed in section (iii).

In section (iv), more complex models of nationalism are synthesized, by drawing together ideas from the literature on this topic and ideas from social psychology. The future possibilities of these models, in verbal and quantitative analysis, are sketched out briefly. These models attained high levels of "statistical significance," and led to noticeable improvements in long-term prediction, in empirical tests described in section (v); however, these tests, based on classical estimation routines, are regarded as

preliminary. The communications concepts of section (iv), as applied in section (v), also yielded an explanation of one of the inconsistencies observed with "gravity models" in previous research; this explanation was validated empirically.

The MIT Cambridge Project has begun implementing the algorithms of Chapter (ii).

(I) GENERAL INTRODUCTION AND SUMMARY

The original purpose of this research was to apply the classical techniques of statistics and econometrics to the Deutsch-Solow model of nationalism, in order to turn this model into a workable tool for predicting the political future. In the course of this research, it became clearer and clearer that the usual statistical methods do a poor job in fitting dynamic models to real-world data, if we judge these models by their ability to make good predictions across time. Furthermore, it became clear that newer and better methods would not be feasible, economically, unless we could also develop new, less expensive algorithms. Thus the goals of this thesis have been five-fold: (i) to describe new ways of fitting models to data; (ii) to define the new algorithms which make these methods feasible; (iii) to introduce evidence for the superiority of the methods (see Table IV-1, on Page IV-34, and the graphs which start on Page VI-3); (iv) to discuss the applications of these ideas, in broad terms, to social and even biological sciences; (v) to discuss the new empirical work on nationalism which has led us in these directions.

Let us begin by discussing the first three goals.

Strictly speaking, we have studied not one, but two, new approaches to fitting models to data, in political science. The first approach was essentially an extension of work by Box and Jenkins, on "ARMA" processes, on "mixed AutoRegressive Moving-Average processes." Chapter (III) discusses the mathematical statistics of this approach, in detail. It begins by pointing out that an ordinary, multivariate autoregressive process, observed by way of data which were not measured perfectly accurately (i.e. measured with random measurement errors or conceptual distortions), turns into a "vector mixed autoregressive moving-average process." It then proceeds to show how the algorithm of "dynamic feedback," discussed in Chapter (II), can be applied to estimate the coefficients of such a process, at a lower cost than was possible with previous methods; the resulting procedure has been tested, and made available to the general user, as part of the MIT Cambridge Project Consistent System. In Chapter (V), we discuss why this approach seemed important to us, in quantitative political science.

In studies of simulated data, this approach did quite a bit better than the best form of regression. In Table IV-1, on Page IV-34, one can see that the

average estimates ("av") produced by "arma," for the coefficients of a simple model, were much closer to the true values than were those of "reg," in the twelve cases studied; this implies much less systematic bias. Also, the dispersion of the "arma" estimates was about half as much as that of "reg," on the whole; this implies less random error in estimation, or, in other words, greater practical efficiency in making use of limited data. However, in studies of real-world data, the long-term predictions by this method were only slightly better than those by regression; for example, in Figures VI-1 through VI-4, on Pages VI-4 through VI-7, one can see that the error distribution curve for "ARMA" is only about 10% smaller in area than that for "Regression," and that the curves are uniformly close to each other.

After re-examining the empirical results of this research, and the concepts of maximum likelihood themselves, we have arrived at a new, more radical and more successful approach to the fitting of models. In essence, the idea is to maximize long-term predictive power <u>directly</u>, over the known data set, instead of maximizing formal likelihood. Formally, this idea rests upon the apriori expectation that many social processes are governed by relatively deterministic underlying

trends, obscured both by measurement noise and by transient deviations of a very complex sort. The qualitative, political basis of this idea is discussed in Chapter (V). The statistical basis is discussed in sections (vii) and (xi) of Chapter (II).

The "measurement-noise-only" approach performed much better than both ARMA and regression, on both real-world and simulated data. In Table IV-1, "ext" is markedly superior to "arma" in estimating coefficients; in the text of Chapter (IV), we note that this superiority is greatest for the simulated data generated by the more complex processes (11 and 12), processes which may be more representative of the real world. In Figures VI-1 through VI-4, the "measurement-noise-only" approach, described as the "robust" approach, had much lower distributions of error than ARMA or Regression did, in long-term prediction. If one allows for the spread of the vertical axis in these graphs, one can see that the "Robust" method cuts the long-term prediction errors in half, roughly; the biggest reductions occur with those variables, such as national assimilation, and with those cases, near the middle of the distributions, for which the simple models of the first half of Chapter (VI) can do an adequate job of prediction.

The empirical results for the "robust" method were all based on special computer programs, designed to take advantage of the simplicity of the models under study. In Chapter (II), we discuss how the general algorithm of "dynamic feedback" can be used to estimate such "measurement-noise-only" models, at a minimal cost, especially for models which may be very intricate, nonlinear and nonMarkhovian; we also discuss how the algorithm can be used to fit more conventional models, to optimize policy, and to perform "pattern analysis" - a dynamic alternative to factor analysis. The technique of "dynamic feedback" is essentially a technique for calculating derivatives inexpensively, to be used with the classic method of steepest descent. In section (iv) of Chapter (III), and in section (iii) of Chapter (VI), we point out how practical experience with steepest descent in this context has led us to new ways of adjusting the "arbitrary convergence weights" of steepest descent; these methods appear to have the power, in normal, practical situations, to speed up the process of convergence by a large factor. In the Appendix to Chapter (II), we mention a few generalizations of these methods, which may be helpful in the general, nonlinear case.

From a practical point of view, the applications of these and other mathematical approaches in the social sciences remain a subject of dispute. The extreme positions of "behaviorism" and "traditionalism" remain popular; a division still tends to exist between quantitative and verbal studies of social behavior. In Chapter (V), we describe the way that our mathematical tools might fit in, in the broader context of social studies and political decision-making. From a utilitarian and Bayesian point of view, we suggest a methodological approach intermediate between "behaviorism" and "traditionalism," in which the different frameworks might be integrated more closely with each other. In sketching out what such an integrated framework might look like, we also point out that the algorithms of Chapter (II), taken as part of "cybernetics," may have some direct value as paradigms, to help us try to understand the requirements of the complex information-processing problems faced by human societies and by human brains. We also mention the possibility of applying these approaches to other fields, such as ecology.

Finally, in Chapter (VI), a few substantive conclusions emerge from our empirical and analytic work on nationalism. The relative success of our long-term

prediction of national assimilation and mobilization, as shown in Tables VI-8 and VI-9 in section (ii) of Chapter (VI), is of some substantive interest; note, for example, that in predicting the percentage of population assimilated, over periods of time on the order of thirty to forty years, that the errors are uniformly distributed between 0% and 2%, except for four outliers (20% of all cases) at 2.68%, 3.08%, 3.09% and 6.21%. The exact sources of weakness in these predictions are also of interest, insofar as they give us a picture of those external factors which really do have the power to divert the processes of assimilation and mobilization from a steady course. In Tables VI-21 and VI-22, in section (iii) of Chapter (VI), we have listed the predictions of the "robust" method for the years 1980, 1990 and 2000; these predictions are subject to caveats discussed in the text of that section. Both in sections (ii) and (iii), all predictions are based on the Deutsch-Solow model, with minor modifications.

In section (iv), more complex models of national assimilation and mobilization are synthesized, by drawing together ideas from the literature on this topic and ideas from social psychology. The future possibilities of these models, in verbal and

quantitative analysis, are sketched out briefly. In section (v), a preliminary test of the models is described. The main methodological conclusion of Chapter (VI) is that the available tools for time-series analysis cannot cope adequately with the level of complexity represented by such models; however, in the preliminary tests, the models attained a high level of "statistical significance," and did have a noticeable value in improving long-term prediction. The communications concepts of section (iv), as applied in section (v), also had the spinoff of suggesting a rational explanation of one of the inconsistencies observed with "gravity models," in previous research; this explanation was validated empirically.

In concluding this introduction, it would seem appropriate to describe what might come out of this work, in the future. However, these possibilities are discussed in enough detail in each of the separate chapters. Still, it should be of general interest that the programming of the general algorithm of Chapter (II) is already underway at MIT, as part of the large-scale "DATATRAN" project on Multics, and is scheduled to be available to the social scientist in 1974.

(II) DYNAMIC FEEDBACK, STATISTICAL ESTIMATION AND SYSTEMS OPTIMIZATION: THE GENERAL TECHNIQUES

(i) INTRODUCTION

In recent years, social scientists and ecologists have become interested more and more in the use of mathematical models to describe the dynamic laws of the systems they study. Karl Deutsch and Robert Solow, for example, have proposed(1) the following model to predict the size of the assimilated population, A, and the unassimilated population, U, in a bilingual or bicultural society:

$$\frac{dA}{dt} = aA + bU$$

$$\frac{dU}{dt} = cU, \qquad (2.1)$$

where "a", "b" and "c" may be treated as constants, at least for medium lengths of time. The constant "b" represents the rate of assimilation, as a fraction of the people yet to be assimilated, per unit of time; "a" and "c" represent the natural growth rates of the assimilated and unassimilated populations, respectively.

Mathematical models may serve two general purposes in the social sciences. On the one hand, they may be used as a tool in verbal reasoning, as a technique for formulating one's assumptions and their consequences very clearly and very coherently; they may be used to construct paradigms, which, like metaphors,

may be very useful but which are not meant to be taken literally. The "prisoner's dilemna" paradigm(2) is a good example of such a model. On the other hand, mathematical models may be used to make actual predictions of variables which can actually be measured; economists, for example, have long been in the business of predicting the GNP, as a number, from the use of equations(3) originated by Keynes and Samuelson. These equations offer different predictions for the GNP, depending on one's assumptions about government spending and tax rates; thus they can be used, not merely in prediction, but in helping the government to choose a policy for spending and taxation which will maximize the real GNP.

Our major concern in this thesis is with the second type of model - predictive models, like the Deutsch-Solow model above. Given such a model, the social scientist would want to ask three questions: (i) how likely is it that the model is true, empirically? ; (ii) how can we measure the values of the constants in the model? ; (iii) if the model is true, but if certain policy-makers could change some of the constants or even control some of the variables directly, what should they do in order to get the "best" results? The first two questions concern the problem of estimation, the core of classical statistics. The third question falls roughly into the area now called "control theory". All three questions can be answered, by use of existing methods, but only for certain restricted classes of models. Our main objective in this chapter is to present a more general

method, to allow us to answer these questions for any explicit model, of any complexity, at a minimal cost in terms of computer time. More precisely, if a user specifies his model in terms of equations built up out of elementary operations and functions, known to a standard computer package, then our method could give this computer package the power to answer the three questions above, at a minimal cost. As more data become available in the social sciences and in ecology, and as models are developed which reflect the true complexity of the social systems themselves, the need for such a general method may grow greater and greater.

In this chapter, we plan to explain the dynamic feedback method, by building up examples of its most important applications; these examples will grow in complexity until, in section (xii), we present the general algorithm explicitly. Thus we will start out in section (iii) by showing how to reduce the cost of conventional nonlinear estimation. In section (iv), we will show how the dynamic feedback method can cope with simple models with "memory"; even simple models of this type are difficult to handle by other methods. In sections (v) through (vii), we will discuss the basic problem of induction, as seen by the statistician. This material prepares us for the discussion of more advanced applications in later sections and also in Chapter (III). In particular, in section (vii), we will propose a new, "robust" approach to estimation, which, <u>even for</u>

simple models, calls for the use of the dynamic feedback method; later, in section (xi), we will specify exactly which "models with memory" are used in this approach, and in Chapters (IV) through (VI) we will discuss the evidence that this approach is worthwhile. In section (xiii), we will discuss the problem of estimation with complex noise models. In section (ix), we will discuss a radically new concept, "pattern analysis," for dealing with situations where the nonlinearities and complexities of a process defy the use of straightforward estimation; the applications of this concept would include problems now dealt with by factor analysis or by pattern recognition techniques. Once a person has finished estimating a model, he may then wish to go on to use this model in formulating policy; in section (x), we will show how the dynamic feedback method can be used at that stage, too, to help one maximize the utility function of one's choice. Finally, in the Appendix, we will mention a few technical procedures, which can help speed up the convergence of a computer routine based on the dynamic feedback method.

(ii) ORDINARY REGRESSION

Let us begin by discussing the first two questions listed on page II-2, from the viewpoint of classical maximum likelihood theory.

How would we ascertain the "truth" of a model like the Deutsch-Solow model, equations (2.1), if we were given the values

of "A" and "U" every year for some nation, from 1901 to 1973? If we were given the values of the constants, a, b, and c, then we could simply solve these equations, starting from the known values of A and U in 1901. In order to avoid having to solve a differential equation, we could rewrite the model in a simpler, but equivalent, form:

$$A(t+1) = k_1 A(t) + k_2 U(t)$$
$$U(t+1) = k_3 U(t), \qquad (2.2)$$

where "U(t)" means the value of U in the year t, and where k_1, k_2 and k_3 are all constants. In either case, we could predict A and U for 1902 through 1973, by starting from our knowledge of A and U in 1901, and using our model. We could compare the predictions of the model against the observed data. And we would discover that equations (2.2) are simply false, as written; there would always be some difference between our predictions and the data, while the equations (2.2) do not allow for any such error. Equations (2.2) are completely deterministic. This complaint may seem like quibbling, but it is central to the classic concepts of statistics. In practice, admittedly, one may be more interested in the <u>predictive power</u> of a simplified model, rather than its formal statistical truth; however, in section (vii), we will be able to discuss this possibility as an extension of the more classical approach discussed here. At any rate, to construct a model which has some hope of being "true", in the

social sciences, we need to express the idea that there will be a certain amount of unpredictable random noise in the system we are studying. Thus we might rewrite equations (2.2) to get:

$$A(t+1) = k_1 A(t) + k_2 U(t) + b(t) \qquad (2.3a)$$

$$U(t+1) = k_3 U(t) + c(t), \qquad (2.3b)$$

where b(t) and c(t) are random error terms, obeying:

$$p(b) = \frac{1}{\sqrt{2\pi}B} \; e^{-\frac{1}{2}(\frac{b}{B})^2}$$

$$p(c) = \frac{1}{\sqrt{2\pi}C} \; e^{-\frac{1}{2}(\frac{c}{C})^2} \qquad (2.4)$$

In other words, we do not know what b(t) and c(t) will be in advance; the probability that b(t) will equal some particular value, b, is given by p(b) in the formula. Strictly speaking, since "b" is a continuous variable, p(b) is actually a probability <u>density</u> function; one may think of it as the probability that b(t) lies between "b", and a nearby point, "b+db", <u>divided</u> by the size of the interval(4), db. These functions for the probability of b and c are simply the classic bell-shaped curve, or "normal distribution." The constants in front of the "e" are there, to make sure that the probabilities of the different values for b add up to one, when the formula is integrated. The constants "B" and "C", like the constants "k_1", "k_2" and "k_3", need to be specified before our model is complete.

According to this elementary model, the probability of b (or c) is highest when b (or c) is zero; in other words, it is highest when the exponent is zero, instead of a negative number. When b gets to be a large number, positive or negative, in proportion to B, the exponent gets to be a large negative number, and the probability falls off very quickly. It should be emphasized that this simple model of noise, while standard, is far from the only possibility in this case; in section (vii), we will mention a few other possibilities.

Once we have decided to formulate such a simple model, at least to start with, classical statistics can tell us exactly how to measure its "likelihood of truth" for any combination of the constants k_1, k_2 and k_3. In sections (v) and (vi), we will discuss in more detail how it is possible for some statisticians to arrive at such strong statements; for the moment, however, we will relegate the theoretical abstractions to a footnote(5). Even on a very concrete level, one can get a feeling for the power of the classical approach.

Looking back at equations (2.3), we may define:

$$\hat{A}(t+1) = k_1 A(t) + k_2 U(t)$$

$$\hat{U}(t+1) = k_3 U(t)$$

"$\hat{A}(t+1)$" is simply the best prediction one could make for $A(t+1)$,

at time t, given our knowledge of A(t) and U(t), and given our model. From equations (2.3), we get:

$$b(t) = A(t+1) - \hat{A}(t+1)$$
$$c(t) = U(t+1) - \hat{U}(t+1). \qquad (2.5)$$

Intuitively, one would expect that a model which gives us "good" predictions, $\hat{A}$, would be likelier to be true than a model which gives us "bad" predictions; one would expect that bigger errors, b and c, would imply a lower probability that the model is true. Indeed, when we look back at equations (2.4), the only probability functions we have with this model, we can see that larger values of b and c would imply a lower probability. More exactly, as we look at these equations, we can see that the probabilities of these errors really depend upon $(\frac{b}{B})^2$ and $(\frac{c}{C})^2$ - i.e the size of the <u>square</u> of the error. As part of our model, we assume that the errors at different times are all independent of each other. Thus in order to combine all the different probabilities for different times, t, into one overall probability, it is legitimate to multiply them all together; this has the effect of telling us to add up all the exponents, the square error terms, $(\frac{b}{B})^2$ and $(\frac{c}{C})^2$, to get an overall measure of the probability of the model. Therefore we can measure the total effective size of the errors in equation (2.3a) by:

$$L_1 = \sum_t (\frac{b(t)}{B})^2 = \frac{1}{B^2} \sum_t (b(t))^2$$

In order to pick the best values of k_1 and k_2, in our model, we do not have to account for the other part of the error, the c^2 term, since our choice of k_1 and k_2 does not affect equation (2.3b). Indeed, to pick the best values of k_1 and k_2, in equation (2.3a), we do not even have to worry about the value of B, since B does not appear in that equation; thus we can simply try to minimize:

$$L = \sum_t (b(t))^2 \qquad (2.6)$$

Similarly, in equation (2.3b), we can pick k_3 by minimizing the analogous function:

$$L' = \sum_t (c(t))^2$$

Notice that we now seem to have two <u>separate</u> measures of truth, for (2.3a) and (2.3b) treated as independent equations. Formally speaking, we have found that the maximization of likelihood for the composite model, (2.3) and (2.4), can be decomposed into the maximization of likelihood for (2.3a) and (2.3b) as separate equations, attached to the top and bottom equations of (2.4), respectively.

This decomposition is due to the simplicity of the original model; it would <u>not</u> be valid for many more complex models. In section (vi), we will present more details of this decomposition with ordinary regression; in Chapter (III), however, we will focus on a class of standard statistical models, the "vector ARMA" models, for which

such a decomposition is impossible, and for which an equation-by-equation estimation procedure cannot have statistical consistency.

Even in the simple case here, however, we have yet to specify how to pick k_1 and k_2 to minimize "L", in equation (2.6). Let us begin by substituting into (2.6) the value of b(t) from equation (2.3a):

$$L = \sum_t (A(t+1) - k_1 A(t) - k_2 U(t))^2 \qquad (2.7)$$

Our problem, again, is to minimize L as a function of k_1 and k_2, while treating the measured data series, A(t) and U(t), as fixed. From basic calculus, we know that a function has its minimum, for variables k_1 and k_2, only at a point where its derivatives with respect to k_1 and k_2 both equal zero. In other words, if the derivative of L with respect to k_1 were <u>not</u> zero, but, say, +10, this means that L will change whenever we change k_1, and that the change in L will equal 10 times the change in k_1, roughly, for small changes in k_1; thus, if we change k_1 by -1/100, then L would change by about -1/10, proving that it hadn't yet reached a minimum at our original choice of k_1 and k_2. Thus we can try to find values for k_1 and k_2 such that the derivatives of L with respect to both of these parameters will equal zero.

Differentiating, we get:

$$\frac{\partial L}{\partial k_1} = \sum_t \frac{\partial}{\partial k_1} (A(t+1) - k_1 A(t) - k_2 U(t))^2$$

$$= \sum_t 2(A(t+1) - k_1 A(t) - k_2 U(t)) \frac{\partial}{\partial k_1}(A(t+1) - k_1 A(t) - k_2 U(t))$$

$$= \sum_t 2(A(t+1) - k_1 A(t) - k_2 U(t))(-A(t))$$

$$= -2\left(\sum_t (A(t+1)A(t) - k_1 (A(t))^2 - k_2 U(t)A(t)) \right),$$

which we will try to set to zero. And we get a similar expression for $\frac{\partial L}{\partial k_2}$; putting them together, we get two algebraic equations:

$$\sum_t A(t+1)A(t) = k_1 \sum_t (A(t))^2 + k_2 \sum_t U(t)A(t)$$

$$\sum_t A(t+1)U(t) = k_1 \sum_t U(t)A(t) + k_2 \sum_t (U(t))^2$$

We can calculate these sums by looking at our data; we can solve these simple simultaneous equations for the variables k_1 and k_2 exactly, by classical algebra, or by using programs available on any computer. The procedure above is the procedure of classic multiple regression.

All of this reasoning, however, supposes that we decide to look at a very simple model, like equation (2.3a). It also assumes that the "errors", b and c, follow a normal distribution. There is nothing to stop us from using the same calculating procedure in cases where

we do not expect the noise to be normal; from a classical point of view, this may still be equivalent to accepting the normal distribution as part of one's "simplified model," but the effects of such a "simplification" are far from obvious apriori.(6).

What happens, however, if we move on to consider a more complex model? What would happen if we decided to change equation (2.3a) itself? For example, in equation (2.3a), we assume that the rate of assimilation, k_2, is constant in any given country. In reality, we know that this is unreasonable. If the "unassimilated" outnumber the "assimilated" by a large majority, they may feel very little pressure at all to assimilate; on the other hand, if they are a tiny isolated group, dependent on an economic world which is mostly "assimilated", then their rate of assimilation is likely to be higher than it otherwise would be. There are other factors involved, but, holding those factors constant, our model is likely to be "truer" and better if it accounts somehow for the power of percentage dominance.

How could we revise equation (2.3a) to express this kind of effect? First of all, we need to find some kind of measure of "percentage dominance." The simplest and most obvious measure is simply the difference between the percentage of the population which is assimilated and the percentage of the population which is unassimilated. In order to avoid having to multiply everything by 100, let us look instead at the difference between the fraction

which is assimilated and the fraction which is not. The fraction of the population assimilated equals, by definition, the ratio between the number of people assimilated, A(t), and the total number of people, A(t)+U(t); thus it equals A(t)/(A(t)+U(t)). The fraction unassimilated equals U(t)/A(t)+U(t); the difference between the two equals (A(t)-U(t))/(A(t)+U(t)). Somehow, we wish to express the idea that an unassimilated person is more likely to assimilate if the "percentage dominance" of the assimilated population is larger. If we recall that k_2 was defined as the rate of assimilation per unassimilated population per unit of time, we may simply postulate that k_2, instead of being constant, will be larger if "percentage dominance", as defined above, is larger. For simplicity, we may consider the idea that k_2 is directly proportional to percentage dominance:

$$k_2(t) = k_2' \frac{A(t)-U(t)}{A(t)+U(t)} \qquad (2.8)$$

This time, k_2' is assumed to be constant. While the actual relation between k_2 and percentage dominance is not likely to be quite this simple, this equation still gives us some expression of the important qualitative idea that there is a strong and consistent positive connection between the two. To generate an <u>explicit</u> model of assimilation, we may substitute this equation back into (2.3a):

$$A(t+1) = k_1 A(t) + k_2' \left(\frac{A(t)-U(t)}{A(t)+U(t)}\right) U(t) + b(t) \qquad (2.9)$$

The second term on the right is an "interaction term," nonlinear in A and U. A great deal of fuss has been made about this kind of nonlinearity, with terminology such as "curvilinear regression", "polynomial regression", and even "spectral regression" often used.(7). However, this kind of situation is fairly easy to deal with. We can solve for b(t), as before, to get:

$$L = \sum_t (A(t+1) - k_1 A(t) - k_2'(\frac{A(t)-U(t)}{A(t)+U(t)})U(t))^2 \qquad (2.10)$$

As a function of k_1 and k_2', this is really the same kind of expression as (2.7), with "U(t)" replaced by a more complicated expression which we might call "U'(t)":

$$U'(t) = \frac{A(t)-U(t)}{A(t)+U(t)} U(t)$$

The derivatives with respect to k_1 and k_2 are the same, with a few "prime" signs interjected, and we wind up with the same algebraic equations to solve and almost the same sums to calculate. (We have to sum up $(U'(t))^2$ and $U'(t)A(t)$ instead of $U(t)^2$ and $U(t)A(t)$.) In practice, one would normally begin by calculating the variable "U'" from one's existing data, and injecting it into a standard regression package to calculate the sums and solve the equations; in a computer package such as TSP, one could compute U' from one's previous data by use of the command "GENR"(generate), and issue a regression command (OLSQ) with U' and A as independent

variables. What is essential in this example is that we continue to express A(t+1) as a linear combination of other variables, which are defined as specific functions of the available data.

(iii) NONLINEAR REGRESSION AND DYNAMIC FEEDBACK

However, if we want to move on to more interesting models of social phenomena, we will often find that we have to estimate constants which do not simply multiply an expression we already know how to calculate, like U'(t); we will find that there are constants on the "inside" of the model. For example, in equation (2.8) we said that k_2 is directly proportional to the dominance of A over U as a fraction of the total population. How do we know that it is a matter of direct proportionality? k_2 is the rate of assimilation, per unassimilated person per unit of time, as originally defined in equation (2.3). In equation (2.8), if U is 25% of the population, then $\frac{A-U}{A+U}$ will equal $\frac{1}{2}$; if U is almost 0, then $\frac{A-U}{A+U}$ will equal 1. Thus we assume that the rate of assimilation will always be twice as much in the latter case, as compared with the former case. But how do we know it is only twice as much? It might be four times as much. After all, the pressures on a tiny community, near 0%, may be much, much larger than on a community near 25%, which may be large enough to protect its own members, and to give them economic

opportunities almost as great as they would find after assimilation. So instead of $\frac{A-U}{A+U}$, we might have written $(\frac{A-U}{A+U})^2$. Or $(\frac{A-U}{A+U})^3$. Even without considering more complicated possibilities, it would be interesting to try to measure just how strong these effects are that we have been talking about; we may write:

$$k_2(t) = k_2' \left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4},$$

where k_4, like k_2', is a constant we would like to estimate. To turn this into an explicit model of assimilation, we substitute into (2.3a):

$$A(t+1) = k_1 A(t) + k_2' \left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4} U(t) + b(t) \quad (2.11)$$

We can solve for b(t), to get:

$$L = \sum_t (b(t))^2$$

$$= \sum_t \left(A(t+1) - k_1 A(t) - k_2'\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4} U(t)\right)^2 \quad (2.12)$$

When we differentiate L, and try to set the derivatives to zero, as before, we find a very unpleasant set of equations emerging:

$$A(t)A(t+1) = k_1 \sum_t (A(t))^2 + k_2' \sum_t A(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4} U(t)$$

$$U(t)A(t+1) = k_1 \sum_t A(t)U(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4}$$

$$+ k_2 \sum_t \left(U(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4}\right)^2$$

$$0 = \sum_t \Big(2(A(t+1) - k_1 A(t) - k_2'\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4} U(t)) \\ * \left(k_2' U(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4} \log \frac{A(t)-U(t)}{A(t)+U(t)} \right) \Big)$$

(Note that we use the asterisk to indicate multiplication, as in FORTRAN.) To solve these three equations as functions of k_1, k_2' and k_4 is not only a difficult exercise in algebra; it would appear to be impossible. There are many equations in algebra for which there simply exist no "closed" solutions - no solutions which can be expressed in terms of the ordinary "vocabulary" of mathematics.(8). Thus, in order to devise computer routines to handle this contingency, we must use routines which give numerical <u>approximations</u> to the constants k_1, k_2' and k_4; we must estimate k_1, k_2' and k_4 by a numerical technique of successive approximations, rather than an exact solution. This is the classic problem of "nonlinear estimation." A similar problem can even arise when dealing with more sophisticated linear models.

There are two well-known methods for dealing with the problem of nonlinear regression. The simplest, and perhaps the best, is the method of "steepest descent." (9). When we try to maximize L, as a function of k_1, k_2' and k_4, we may not be able to solve for $\frac{\partial L}{\partial k_1} = 0$, $\frac{\partial L}{\partial k_2'} = 0$ and $\frac{\partial L}{\partial k_4} = 0$. However, if we start off with reasonable guesses for all of the constants, k_1, k_2' and k_4, then we can <u>differentiate</u> (2.12), plug in our guesses, and see if the

derivatives happen to equal zero; if so, chances are good that we have guessed the best values. (With classic regression, when we had two simple equations in two unknowns, k_1 and k_2, there would only be one solution in the usual case, and there would always be a minimum for L; therefore, we could be fairly certain that the derivatives were zero only at the minimum. With very complex formulas, we simply have no way of being sure about this.)

If the derivatives are not zero, then we can guess new values for our constants, values which will make L smaller. If $\frac{\partial L}{\partial k_1}$ is positive, then we can decrease L by decreasing k_1; if $\frac{\partial L}{\partial k_1}$ is negative, then we can decrease L by <u>increasing</u> k_1. If $\frac{\partial L}{\partial k_1}$ is close to zero, then k_1 is probably close to its best value; if $\frac{\partial L}{\partial k_1}$ is far away from zero, then k_1 is probably further off. Thus we can create a new guess, $k_1(n+1)$, better than the old guess, $k_1(n)$, by changing k_1 in <u>proportion</u> to $\frac{\partial L}{\partial k_1}$, but in the <u>opposite</u> direction:

$$k_1(n+1) = k_1(n) - C\frac{\partial L}{\partial k_1} ,$$

where C is some positive constant, and where we calculate the derivative by using our <u>old</u> guesses, $k_1(n)$, $k_2'(n)$ and $k_4(n)$. We also calculate the derivative and then the new guess for k_2' and k_4, each. Once we have our new guesses for k_1, k_2' and k_4, we can go back to (2.12), to see if we really have gotten a smaller value for L. If we have, then we can start again from our new guesses, to check the derivatives, etc. If <u>not</u>... then C must be

too big. If C is small enough, the definition of the derivative assures us that L can be predicted as well as we like by looking at just the first derivative; therefore, for some C small enough, we know that our new guesses will have a smaller L than our old guesses. If we find ourselves making C smaller and smaller, from guess to guess, then we may eventually quit, when C is so small that we aren't changing the constants very much. Hopefully, this will mean that our approximations are very close to the ideal values. In the Appendix to this chapter, we will suggest a few ways to speed up the convergence of this classical technique.

As we look back at equation (2.12), it is clear what our biggest problem is in actually doing all this work: we have to calculate the derivatives of a very complicated-looking expression, L, and we have to calculate the exact numerical values of these derivatives for many different values of the constants k_1, k_2' and k_4. Even worse, there is the question of who the "we" is who will do all the work, in most cases. Classical regression can be done automatically for the social scientist, at a low cost, by a computer program; the social scientist need only load in his data, and specify his choice of variables. Who is to do the differentiating here? The social scientist? In BMD, one of the biggest computer packages for use in social science and biology, there is only one "nonlinear regression" routine, added into the X-Supplement(10) available June 1972;

this routine requires the social scientist to write his own FORTRAN programs both for the function A(t+1) and for all its derivatives. It is reasonable to ask a social scientist to understand the logic behind a formula like (2.11); it seems rather unreasonable to ask him to carry out elaborate differentiations, and write and debug his own FORTRAN programs, for every such model he chooses to investigate. Also, this approach could become expensive in terms of computer time, too, depending on the user's ability to devise low-cost ways of calculating his derivatives. There is a second possibility: the user could be asked to specify a FORTRAN program to calculate A(t+1), as a function of A(t), U(t), k_1, k_2' and k_4; the program would then go on to calculate the derivatives numerically, by changing k_1 a little bit, and seeing what happens to A(t+1). At each time, for each constant, the computer would have to carry out calculations as expensive as calculating A(t+1); with many constants, this could multiply the cost many-fold. In the two nonlinear regression routines easily available in Cambridge, besides BMD - TSP-CSP(11) and Troll/1(12) - the social scientist has a more convenient way to get his work done. In these two systems, he need only specify his model in terms of a "formula", like:

```
A(t+1) = k1*A(t) + k2*U(t)*(((A(t)-U(t))/(A(t)+U(t)))**k4)
```

This is the same as (2.11), but with FORTRAN conventions used to make it possible to put everything on one line, and with error terms

Actual Variable	Variable Number (Address)	Category	Major Source	Minor Source
$A(t+1)$	13	sum	12	11
$k_1 A(t)$	12	product	3	1
$k_2 U(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4}$	11	product	10	4
$U(t)\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4}$	10	product	9	2
$\left(\frac{A(t)-U(t)}{A(t)+U(t)}\right)^{k_4}$	9	power	8	5
$\frac{A(t)-U(t)}{A(t)+U(t)}$	8	ratio	7	6
$A(t)-U(t)$	7	difference	1	2
$A(t)+U(t)$	6	sum	1	2
k_4	5	parameter	-	-
k_2	4	parameter	-	-
k_1	3	parameter	-	-
$U(t)$	2	given	-	-
$A(t)$	1	given	-	-

Table II-1: Table of Operations for Equation (2.11)

left implicit. (Single asterisk means multiplication, double means raising to a power.) This formula then gets translated by the computer into a list of simple expressions. This list of expressions would normally look something like Table II-1, with the explanation column on the left removed; such a list is called a "Polish string." The categories of operation allowed on such a list depend on the arbitrary choices of the systems programmers. In some systems, there are function names reserved for user-supplied FORTRAN subroutines; in other systems, there are functions corresponding to model neurons, for use in statistical pattern recognition; et cetera. It is already possible for a computer to calculate the symbolic derivatives of a formula by manipulating formulas which have been broken down like this; however, that process becomes quite expensive, if we have many parameters to differentiate against.

The easiest way to calculate these derivatives is by a simple use of dynamic feedback. Now we know that:

$$L = \sum_t (b(t))^2$$

$$\frac{\partial L}{\partial k_1} = \sum_t \frac{\partial}{\partial k_1}((b(t))^2)$$

To calculate $\frac{\partial L}{\partial k_1}$, we need only calculate $\frac{\partial}{\partial k_1}((b(t))^2)$ for <u>each</u> time, and add up these derivatives over time. We want to know the effect on the error, $(b(t))^2$, of <u>changing</u>, say, k_1, while we keep our data

(i.e. A(t),A(t+1),U(t)) constant, and while we keep the other parameters (k_2',k_4) constant. In Table II-1, let us define the "ordered derivative" of $(b(t))^2$ with respect to variable number i to be the change we get in $(b(t))^2$ in proportion to the change in variable #i, when we hold all the previous variables constant. For $\frac{\partial}{\partial k_1}(b(t))^2$, this definition doesn't give us anything new; the ordered derivative, $\frac{\partial^+}{\partial k_1}(b(t))^2$, is the same as the ordinary partial derivative, $\frac{\partial}{\partial k_1}(b(t))^2$. But for the other variables, it gives us something new to calculate.

Now: suppose we ask, in changing variable #7 by a small amount, what will the total effect be on $(b(t))^2$? Changing variable #7, we will have a direct effect on only one variable, later in the system, variable #8. (See Table II-1.) Thus if:

$$X_7' = X_7 + d,$$

where "d" is a small number, where X_7 is variable #7, and where "X'" is the value after our changes. We will produce the following direct effect on later variables:

$$X_8' = \frac{X_7'}{X_6'} = \frac{X_7 + d}{X_6} = X_8 + \frac{d}{X_6} \qquad (X_6 \text{ held constant.})$$

If we are calculating backwards from the top of the table, we already know $\frac{\partial^+}{\partial X_8}((b(t))^2)$; we already know the ratio between $(b(t)')^2-(b(t))^2$ and $X_8'-X_8$. Let us call that ratio, or derivative,

"S_8". Thus we know, for small values of g, that if $X_8' = X_8 + g$, that $b'^2 = b^2 + S_8 g$. Now we just found out, for small d, that if $X_7' = X_7 + d$, that $X_8' = X_8 + \frac{d}{X_6}$; thus if we write $g = \frac{d}{X_6}$, we know that this change in X_8 will lead to a <u>total</u> change in $(b(t))^2$ of $S_8 \frac{d}{X_6}$. (Before, when we measured S_8, we assumed that X_7 would be held constant. However, when we vary X_7, and hold the earlier variables constant, there is no way that this change can affect anything later on, except by way of X_8.) Thus we deduce:

$$S_7 = \frac{S_8}{X_6}$$

In more sophisticated language, this is an example of:

$$\frac{\partial^+}{\partial X_7} ((b(t))^2) = \left(\frac{\partial^+}{\partial X_8} ((b(t))^2)\right) \frac{\partial f_8}{\partial X_7} ,$$

where f_8 is the function $X_8 = f_8(X_7, X_6) = \frac{X_7}{X_6}$. Now let us consider a more complicated example. In Table II-1, X_2 has a direct effect on <u>three</u> variables higher in the table - X_{10}, X_7 and X_6. When we start to vary X_2, we have to account for the total effect of all <u>three</u> of the changes it introduces directly on these other variables. Thus we get:

$$S_2 = S_{10} \frac{\partial f_{10}}{\partial X_2} + S_7 \cdot \frac{\partial f_7}{\partial X_2} + S_6 \cdot \frac{\partial f_6}{\partial X_2}$$

$$= S_{10} X_9 + S_7(-1) + S_6(+1).$$

Of course, X_2 is simply U(t); the reader, differentiating (2.12) with respect to U(t), would also arrive at three terms, equal to

the three terms here, but the work involved would be rather tedious. To make it explicit how we begin this downwards calculation, let me point out how to get S_{13}:

$$S_{13} = \frac{\partial^+}{\partial \hat{A}(t+1)} ((b(t))^2)$$

$$= \frac{\partial^+}{\partial \hat{A}(t+1)} ((A(t+1) - \hat{A}(t+1))^2)$$

$$= -2(A(t+1) - \hat{A}(t+1))$$

One way to operationalize all this is to start from the top, and, for every variable, look at all of its direct connections to variables higher up. An easier way, in practice is to pass down from the top all the information to variables below them and also directly connected to them; the effect is exactly the same, but the order of computations is easier to deal with. We can start out, in our example, by setting S_1 through S_{12} to zero, and plugging in S_{13} as above. At S_{13}, we note that we have a "sum"; thus we add S_{13} to S_{12} and to S_{11}, to account for the direct effect of X_{12} and of X_{11} on $(b(t))^2$. Then we are done; we go down to S_{12}. At S_{12}, we know that all the later effects of X_{12} have already been added into S_{12}, and that our value for S_{12} has been completely calculated. At S_{12}, we encounter a product, X_3X_1. Thus we add $S_{12}X_3$ to S_1, and $S_{12}X_1$ to S_3. We go down to S_{11}. We encounter another product. We add $S_{11}X_{10}$ to S_4, and $S_{11}X_4$ to S_{10}. We go down to S_{10}. And so on. At the end, we really look only at S_5, S_4 and S_3, the derivatives we wanted for the

steepest descent method. The mathematical basis of these operations is the theorem, for a set of ordered functional relations $X_i = f_i(X_{i-1}, X_{i-2}, \ldots X_1)$, that:

$$\frac{\partial^+ X_n}{\partial X_i} = \sum_{j=i+1}^{n} \frac{\partial^+ X_n}{\partial X_j} \cdot \frac{\partial f_j}{\partial X_i} \quad , \tag{2.13}$$

a theorem to be discussed in section (xii). For each line of the list, as we go down, we have only two calculations to perform at most, one for the "major source" and one for the "minor source"; thus the total number of calculations needed for each time, t, will equal only $2n_x$, where n_x is the total number of variables on the list. The total cost will be $2n_x T$, across all times, to get all of the derivatives we want, regardless of how many parameters there are. Notice that to go up the list, starting from U(t) and A(t), requires one calculation per line of the list; thus the total cost, merely to compute all the $\hat{A}(t+1)$ for a given model, will equal $n_x T$, the same order of magnitude. I assumed, above, that we had already carried out this latter calculation, so that the values of the "X_i" were already known; given that we have to find L for each guess, not just the derivatives of L, there is no extra cost in first calculating the "X_i" and L.

In practice, one can imagine three ways that a systems programmer might want to use the generalized form of the dynamic feedback method. First of all, he might simply write a subroutine,

to do the calculations specified above directly, on the table of operations for some model. Second of all, he might write one subroutine to look at one table of operations, and to specify the calculations required by the dynamic feedback method for this table, in another table of operations; he might write a second subroutine to prune away all unnecessary and redundant operations from this table. The relative advantages of these methods would depend heavily on the characteristics of the model being studied, on the number of time periods and calculations of derivatives to be performed, and even on machine characteristics. Finally, one might imagine the possibility that the operations on a table like Table II-1 will someday be grouped into "strata," groups of operations that can be performed in parallel, on a computer capable of parallel processing. On such a machine, one could perform the operations at a given set of "S_i", in parallel, using the same procedures as above, so long as none of the corresponding "X_i" depend directly on each other as input sources; in short, one could use any system of stratification which was adequate for calculating the X_i. This possibility is restricted, however, by the requirement that several processors would have to be able to add something to the same machine word (S_i for "i" on the next lower stratum), at the same time, with the result that this word would be increased by the sum of all the numbers added.

(iv) MODELS WITH MEMORY

Now: with a firm mathematical basis for these procedures, equation (2.13), we can extend them still further. The models we have discussed so far have all been rather conventional "Markhovian" models; in other words, they give us a prediction of A(t+1) as a function of A(t) and U(t). We could add in A(t-1) and U(t-1) as dependent variables, without changing much, because we would still have a distinct table for every time "t" giving $\hat{A}(t+1)$ as a function of a manageable number of variables. Suppose, however, that we have a model with "memory." In economics, for example, there is a model of consumer behavior which states that consumers spend money, not in proportion to their <u>current</u> income, but in proportion to the <u>permanent income</u>(13) which they expect to average in their lifetimes; the model states that the perceived permanent income is adjusted slightly, from year to year, in response to actual income. Thus we get a model:

$$\begin{aligned} C(t) &= k_1 Y_p(t) + b(t) \\ Y_p(t) &= (1-k_2)Y_p(t-1) + k_2 Y_A(t), \end{aligned} \tag{2.14}$$

where "C" is consumption, "Y_p" is permanent income, "Y_A" is actual annual income, and "b(t)" is an error term. Note that statistics will normally be available here for "C" and for "Y_A", <u>not</u> for Y_p. However, this is still what we would call an "explicit" or

"phenomenological" model. Given estimates for $Y_p(t)$, and data for $Y_A(t)$, the model tells us exactly how to calculate $Y_p(t+1)$ and how to predict C(t). To calculate the $Y_p(t)$, for all times t, and to make predictions for the C(t), we need to start off, at time t=1, with some estimate of $Y_p(0)$; this estimate we can treat as an external constant of the model, like k_1 and k_2, to be estimated by the statistician (us). (From $Y_p(0)$ and $Y_A(1)$, equation (2.14b) tells us how to calculate $Y_p(1)$; then we can calculate $Y_p(2)$ from $Y_p(1)$ and $Y_A(2)$, then $Y_p(3)$ from $Y_p(2)$ and $Y_A(3)$, etc.)

To minimize the sum of the errors squared, L, is much harder in this case than with our complicated-looking model in equation (2.11). To calculate $\frac{\partial L}{\partial k_1}$ here, it is not enough to set up separate tables, like Table II-1, for each time t, and add up the $\frac{\partial}{\partial k_1}(b(t)^2)$. Equation (2.14b) establishes a <u>connection</u> between the unknown variables, Y_p, at all different times. However, we <u>can</u> set up a large table to include <u>all</u> the different values of $Y_p(t)$ and C(t) across different times; this will be like taking the separate tables for each time t, tables like those implied by Table II-1, and putting them together into one large table. In this large table, we can show the relations that exist across time. Suppose that we have data for "C" and "Y_A" from time 1 to time 4. We get a big table, as shown in Table II-2, on the next page.

With a given set of constants - $Y_p(0)$, k_1 and k_2 - and with a

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source(s)
L	37	sum	36	28,20,12
$(b(4))^2$	36	product	35	35
$b(4)=C(4)-k_1Y_p(4)$	35	difference	34	33
$C(4)$	34	input	-	-
$k_1Y_p(4)$	33	product	32	1
$Y_p(4)$ (see 2.14b)	32	sum	31	29
$k_2Y_A(4)$	31	product	30	2
$Y_A(4)$	30	input	-	-
$(1-k_2)Y_p(3)$	29	product	24	4
$(b(3))^2$	28	product	27	27
$b(3)=C(3)-k_1Y_p(3)$	27	difference	26	25
$C(3)$	26	input	-	-
$k_1Y_p(3)$	25	product	24	1
$Y_p(3)$ (see 2.14b)	24	sum	23	21
$k_2Y_A(3)$	23	product	22	2
$Y_A(3)$	22	input	-	-
$(1-k_2)Y_p(2)$	21	product	16	4

Table II-2: Table of Operations for Equations (2.14).
(top section)

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$(b(2))^2$	20	product	19	19
$b(2)=C(2)-k_1Y_p(2)$	19	difference	18	17
$C(2)$	18	input	-	-
$k_1Y_p(2)$	17	product	16	1
$Y_p(2)$	16	sum	15	13
$k_2Y_A(2)$	15	product	14	2
$Y_A(2)$	14	input	-	-
$(1-k_2)Y_p(1)$	13	product	12	4
$(b(1))^2$	12	product	11	11
$b(1)=C(1)-k_1Y_p(1)$	11	difference	10	9
$C(1)$	10	input	-	-
$k_1Y_p(1)$	9	product	8	1

Table II-2: Table of Operations for Equations (2.14). (middle section)

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$Y_p(1)$	8	sum	7	5
$k_2 Y_A(1)$	7	product	6	2
$Y_A(1)$	6	input	-	-
$(1-k_2)Y_p(0)$	5	product	3	4
$1-k_2$	4	difference	0	2
$Y_p(0)$	3	parameter	-	-
k_2	2	parameter	-	-
k_1	1	parameter	-	-
1	0	input	-	-

Table II-2: Table of Operations for Equations (2.14). (bottom section)

given set of data, we can calculate "forwards in time", or upwards in this table, to calculate every one of the "actual variables," including L, the total error. To calculate $\frac{\partial L}{\partial k_1}$, we can calculate backwards, just as we did before with Table II-1, from the top of the table to the bottom. This time, however, it is easier to see where to start:

$$S_{37} = \frac{\partial^+ L}{\partial X_{37}} = \frac{\partial L}{\partial L} = 1$$

This time, with L itself on top, instead of $(b(t))^2$, we get a simpler result at the end:

$$S_1 = \frac{\partial^+ L}{\partial X_1} = \frac{\partial L}{\partial k_1}$$

$$S_2 = \frac{\partial^+ L}{\partial X_2} = \frac{\partial L}{\partial k_2}$$

$$S_3 = \frac{\partial^+ L}{\partial X_3} = \frac{\partial L}{\partial Y_p(0)} \quad ,$$

exactly the quantities we need to apply the steepest descent method. For each line of the table, except for the top, there are only two sources, or no sources; thus to go back from the top line to the bottom line requires only two operations per line, at the most. For a very large table, with n_x lines for each time t, and with T periods of time, this amounts to $n_x T$ lines, and $2n_x T$ operations in all, to get all the derivatives of L. Remember that to go up

the table, to calculate L, we had to carry out one operation per line - $n_x T$ operations in all. No matter how complex the model, if the functional relations across time are explicit enough that they can be put into formulas which the computer can translate into a table, like Table II-2, then "dynamic feedback" can be used to calculate all the derivatives, in one pass.

As a practical matter, one may wonder just how explicit is "explicit enough". In general, the procedure above allows us to calculate the derivatives backwards down any ordered table of operations, so long as the operations correspond to differentiable functions. In order for us to use this method, then, the primary requirement is that we be able to specify the model well enough to construct such a table. This is the <u>same requirement</u> that applies when we wish to use a model forwards in time, to make a prediction of the future, without having to solve a complex set of nonlinear algebraic equations in every time period. In general, in the existing computer packages (including FORTRAN compilers), any formula expressed in the following form can be parsed into a table of operations (a "Polish string") generating the variable $X_i(t)$ from operations performed on the arguments:

$$X_i(t) = f_i(\text{arguments}),$$

where "f_i" is a function made up by nesting basic operations known to the computer package; for example:

$$X_i(t) = W(t-1)*Y(t-1) + k + \sin(Z(t-1)).$$

In order for a set of such formulas to be converted into a table of operations, we need only find an ordering of the variables to be computed, "$X_i(t)$", such that the arguments used in calculating $X_i(t)$ are calculated before $X_i(t)$ itself is; the table of operations to calculate $X_i(t)$ can simply be inserted on top of the table already built up to calculate variables earlier in the causal ordering. If the arguments of "f_i" included only constants, parameters and values of variables at "t-n", for all f_i, with "n" always greater than zero for endogenous variables, then this requirement would be satisfied automatically. Otherwise, an ordering of the variables $X_i(t)$ would have to exist, with the later expressed as functions of the earlier. Global things to be calculated, such as the sum of a utility function or a loss function over time, can always be inserted on top of the table of calculations, so long as we specify formulas for calculating them as a function of sums across time or the like. Indeed, even if one had a set of implicit equations, so that one had to use algebraic solution methods instead of explicit calculation in order to carry out a simple prediction of the future from given parameters, then one could easily calculate the matrix of partial derivatives for those equations, to be used in conjunction with the algebraic solutions generated for prediction, to allow one to carry out dynamic feedback estimate; however, simulations of this sort are both expensive, and outside the major realm of interest here.

Parenthetically, one might note that there is a certain difference between the operations needed to specify the generation of a random number, and the operations needed to calculate the associated loss function. Estimation by dynamic feedback requires the specification of loss functions. In the case where the unobserved random numbers are generated by a rather complex process, the translation between the two forms of specification may not be easy. However, if the losses one is concerned with are the discrepancies between the actual and predicted values of known variables, the specification of an explicit loss function should present no problem to the user of a computer package. The corresponding model would be suitable for predicting the future, but may not be quite as suitable for stochastic simulations of the future, in some cases. In such cases, however, the method of pattern analysis, to be discussed in section (ix), may help reduce the distance between the two forms of specification.

(v) NOISE, AND THE CONCEPT OF THE TRUTH OF A MODEL, IN STATISTICS

Up until now, we have avoided one other aspect of statistical estimation: the problem of noise models. In our old model, in equations (2.3) and (2.4), we assumed a simple equation to predict $A(t+1)$, and a simple bell-shaped curve for the distribution of the

errors, b(t). In the last thirty pages or so, we have considered more and more complex models to predict things. However, we have stayed with the old idea of minimizing the square of the error, an idea based upon the old bell-shaped normal distribution. We have begun in this way only with great reluctance, and only for the sake of exposition. In fact, if we admit that most processes in human society and ecology do contain important elements of randomness, then we must admit that equation (2.4) is just as much a part of our original model as equation (2.3). Equation (2.4) is not an "assumption which must be proven true before we can use classical techniques"; like equation (2.3) itself, it is part of a simple, approximate model, to be <u>evaluated</u> for its predictive power. Unfortunately, there has sometimes been a tendency in social science and ecology to formulate ever more complicated models to predict things, <u>without</u> an explicit model of the random element; the "errors" are sometimes regarded as something unpleasant, that one faces up to at the end of one's research, after one has formulated a model of what is interesting.

Statisticians, on the other hand, have long since passed the stage of "minimizing least squares" or of "minimizing error" in general. We mentioned, earlier, the idea of measuring the "probability of truth" of a model. We mentioned the problem of how to estimate the probability of truth of a model, <u>given</u>

a set of data observations. The traditional "maximum likelihood" school of statistics, as represented by Jeffreys and Carnap(14), and the more recent Bayesian schools, both agree that this is simply a problem in conditional probabilities: how do we estimate the probability of the truth of a model, conditional upon our having made a certain set of observations? Formally, the conditional probability of A given B, $p(A|B)$, is defined to equal $p(A \text{ and } B)/p(B)$; from this follows Bayes' Law:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Statisticians have applied this law to deduce:

$$p(\text{model}|\text{data}) = \frac{p(\text{data}|\text{model})p(\text{model})}{p(\text{data})} \qquad (2.15)$$

This equation does not say that the "data" and the "model" have to be expressed in purely mathematical terms; as a result, the equation has led to enormous controversy both among statisticians and among philosophers. It is a general equation telling us how to determine the probability of truth of any sort of theory; thus, its relevance to social and natural science goes well beyond the question of statistical methods proper. The calculations on the right involve two terms of real interest - $p(\text{data}|\text{model})$ and $p(\text{model})$. The term $p(\text{data})$ is the same for all models, and does not help us to evaluate the relative probabilities of truth of different models, except perhaps indirectly(15). The term $p(\text{data}|\text{model})$ represents what

statisticians have traditionally focused on: how well the data "fit" the model, defined as the probability that these particular data would have been generated if the model were true. The term p(model) refers to the probability of the model before any data have been observed at all; it is our apriori probability distribution. The philosopher Immanuel Kant long ago asserted that "empirical induction" is impossible, without some system of apriori assumptions (the "apriori synthetic") with real information content to them(16); the choice of "p(model)" would constitute such a system of assumptions.

More recent philosophers, such as Carnap and Jeffreys(17), have tried to preserve the more popular attitudes of pure empiricism and positivism, by suggesting that p(model) should be "equal" (apriori) for all different models. Thus p(data|model) would be the only term left to consider, in measuring probabilities of truth. Their suggestions have been carried over to the field of statistics, where they are now orthodox practice(18). This approach is normally referred to as the "maximum likelihood approach." In more recent years, however, many members of a new school of statisticians, the Bayesians, have grown in their opposition to this orthodox procedure. They have pointed out that "p(model)=k", with the same "k" for all models, is a very strong assumption, just as strong as any of the alternatives. In most practical problems, the social scientist would have some reason to expect some models to be likelier than others, even before

he runs his statistical analysis. Thus they suggest that a user of statistical programs should be asked to specify his apriori probability distribution as the first step of any statistical analysis(19); then the computer program can account for both p(model) and p(data model), in picking out the model with the highest probability of truth. From a broader perspective, one might say that the Bayesians are proposing a procedure for allowing the social scientist to account for two different kinds of data - statistical data, and verbal data he has from other sources. This still leaves open the question of where his initial p(model) should come from, a question which we can avoid in this context.(20).

The Bayesians may be right in principle, but in practice the orthodox procedures may remain a sensible way to design computer statistical packages. The social scientist, when he reads the output of a computer, would normally expect that this output reflects only the ability of different models to fit the actual data; in deciding what he finally believes about the world, he can then account for his verbal data. This does require that he understand what "standard errors" mean, in ordinary regression, so that he can get some idea of the variety of models consistent with the statistical data. It also suggests that a direct printout of the relative probabilities of truth of different models, over the given data,

would be a useful feature to have. In brief, it requires the development of an intuition regarding the relation of mathematics to social processes, an intuition strong enough to sustain the balanced assessment of probabilities. This does place a burden on the social scientist. On the other hand, the extreme Bayesian alternative - to ask a social scientist to encode his intuition into a few normal distributions, and to ask for a more complete faith in what comes out of a computer - would seem to place a much heavier burden on the social scientist. It would tend to de-emphasize the learning experience which usually occurs at the end of a statistical analysis, when the social scientist tries to relate all the things which came out of the computer to what he knows in the real world; if this experience is what develops a balanced intuition in the first place, it should not be given a diminished role. In Chapter (V), we will discuss in detail the importance of this type of experience to the actual application of statistical methods in the social sciences. Furthermore, the verbal knowledge of a social scientist will not normally fit a simple distribution. Even if it did, few "intuitive decision-makers" can express their intuition at all reasonably in terms of probability distributions, even in simple cases, without extensive training in that task.(21). While there is more that could be said on both sides of this particular issue, the orthodox approach would seem quite adequate for the purposes of the present context.

The dynamic feedback algorithm, which we are discussing in this paper, can actually be applied to Bayesian estimation as easily as to conventional estimation. In our examples and in our applications we will follow the more orthodox procedures. However, we will refer back, on occasion, to the concept of "prior probabilities" when this is appropriate.

In concluding this section, we might note, for the sake of the mathematician, that equation (2.15), when used in statistics, is normally used to give the probability distributions of a continuous family of models, rather than a discrete probability. Strictly speaking, such distributions are not even functions; they are actually "measures", and they would normally be written as the product of a function times an explicit measure, like "$d\theta$", where "θ" is a parameter of the model. In equation (2.15), however, the choices of measure used for the data cancel out; thus it is not of intrinsic importance, so long as we are consistent in the units we use to record the data. The choice of measure for the model is one more aspect of the problem of specifying p(model), an aspect discussed by the sources referred to above. In general, however, one would not expect the maximum likelihood choice of parameters to be affected very strongly by the choice of measure, unless the standard error for these parameters indicated a large uncertainty and probably a low statistical significance in any case.

(vi) ORDINARY REGRESSION AND THE MAXIMUM LIKELIHOOD APPROACH

Back in section (ii), when we discussed how to measure the "probability of truth" of the Deutsch-Solow model, we glossed over the basic questions discussed in section (v). In section (ii), we found ourselves discussing two different "measures" of the probability of truth, one for equation (2.3a) and another for equation (2.3b); all the rest of our discussion focused on equation (2.3a), an equation to predict a single variable, A(t+1). When discussing a single equation, to predict a single variable from known data, it makes some kind of sense simply to add up all the square errors $(b(t))^2$ across time, and use the sum as a measure of how good the equation is. However, what do we do if there are two equations and two sets of errors? How do we combine the two different error terms to measure the validity of the model as a whole? With the Deutsch-Solow model, we could pick out the best values for k_1 and k_2 without answering these questions, because the two equations were essentially independent, and because we assumed that the two error terms, "b" and "c", each had their own probability distributions independent of each other. (Equation (2.4).) But we had to avoid the question of how to measure the validity of the model as a whole. Now, using the concepts of section (v), we can come back

to answer this question. Let us define the relative probability of truth, P, of any model, as:

$$P = p(\text{data} \mid \text{model}),$$

the probability that we would have observed the data we have observed, if the model were true. More precisely, this "probability" is actually a probability density, as are the other "probabilities" in this section. For simplicity, let us assume, with the Deutsch-Solow model, that we only have data for three years, 1958, 1959 and 1960, in one country. Writing out the data explicitly, we are trying to measure:

$$P = p(A(1960),U(1960),A(1959),U(1959),A(1958),U(1958) \mid \text{model}),$$

which, by classic probability theory, equals the product:

$$\begin{aligned} P = {} & p(A(1960) \mid U(1960),A(1959),U(1959),A(1958),U(1958),\text{model}) \\ & * p(U(1960) \mid A(1959),U(1959),A(1958),U(1958),\text{model}) \\ & * p(A(1959) \mid U(1959),A(1958),U(1958),\text{model}) \\ & * p(U(1959) \mid A(1958),U(1958),\text{model}) \\ & * p(A(1958),U(1958) \mid \text{model}). \end{aligned}$$

Now our model, equation (2.3), predicts A(1960) as a function of A(1959) and U(1959); once these data are given, the other data will not affect the probability of A(1960) as given by the model.

Similarly for U(1960), etc.; thus we can simplify our expression:

$$\begin{aligned} P = {} & p(A(1960) \mid A(1959), U(1959), \text{model}) \\ & * p(U(1960) \mid U(1959), \text{model}) \\ & * p(A(1959) \mid A(1958), U(1958), \text{model}) \\ & * p(U(1959) \mid U(1958), \text{model}) \\ & * p(A(1958), U(1958) \mid \text{model}). \end{aligned}$$

Once we are given values for A(1959) and U(1959), how do we determine the probabilities of the possible values for A(1960)? The most likely value of A(1960), according to equations (2.3) and (2.4), is the one with b(1960)=0, i.e. A(1960) = A(1960), which equals $k_1A(1959)+k_2U(1959)$. But any value for A(1960) would be consistent with equations (2.3), for some value of b. Values of A(1960) far away from A(1960), however, would imply large values of b, which, according to equations (2.4), are not as likely as small values of b. To determine the probability of any given A(1960), given A(1959) and U(1959), we need only look at the probability of the value of b needed to generate the combination, $b(1959)=A(1960)-k_1A(1959)-k_2U(1959)$. Thus:

$$\begin{aligned} p(A(1960) \mid A(1959), U(1959), \text{model}) &= p(b(1959) \mid \text{model}) \\ &= \frac{1}{\sqrt{2\pi}B} \, e^{-\frac{1}{2}\left(\frac{b(1959)}{B}\right)^2} \end{aligned}$$

And similarly for A(1959), U(1960) and U(1959). Thus we get, in summary:

$$P = p(b(1959)\,|\,\text{model}) * p(b(1958)\,|\,\text{model}) * p(c(1959)\,|\,\text{model}) * p(c(1958)\,|\,\text{model}) * p(A(1958),U(1958)\,|\,\text{model}),$$

which, by equation (2.4), equals:

$$\left(\frac{1}{\sqrt{2\pi}B}\, e^{-\frac{1}{2}\left(\frac{b(1959)}{B}\right)^2}\right)\left(\frac{1}{\sqrt{2\pi}B}\, e^{-\frac{1}{2}\left(\frac{b(1958)}{B}\right)^2}\right)\left(\frac{1}{\sqrt{2\pi}C}\, e^{-\frac{1}{2}\left(\frac{c(1959)}{C}\right)^2}\right) * \left(\frac{1}{\sqrt{2\pi}C}\, e^{-\frac{1}{2}\left(\frac{c(1958)}{C}\right)^2}\right) * p(A(1958),U(1958)\,|\,\text{model})$$

What do we do about the last term, representing our earliest data point, 1958? The usual practice is simply to ignore the final term on grounds that it is difficult to compute, and contributes only one time-point worth of information; for long series of data, the importance of one extra point of information grows very small. In the social sciences, the argument for eliminating this term grows even stronger. This term, as usually interpreted(22), requires us to compute the probability that we would have started off at a data point equal to (A(1958),U(1958)), if this initial data had been generated by the Deutsch-Solow model operating for an infinite length of time before the start of the available data. Normally, in the social sciences, one picks the start of one's data series for one of two reasons: (i) one is trying to find a model to describe events in a given historical period, and one does not expect the

model to be valid before the start of one's data series; (ii) the data are not available before a certain time, usually implying that some aspects of the social system were different beforehand. Furthermore, if one's model is not "stationary", as few stationary processes are, then the usual procedure for computing this term breaks down in any case.

On this basis, we get a relative probability density:

$$P = \frac{1}{\sqrt{2\pi}B} e^{-\frac{1}{2}(\frac{b(1959)}{B})^2} \frac{1}{\sqrt{2\pi}B} e^{-\frac{1}{2}(\frac{b(1958)}{B})^2} \frac{1}{\sqrt{2\pi}C} e^{-\frac{1}{2}(\frac{c(1959)}{C})^2} \frac{1}{\sqrt{2\pi}C} e^{-\frac{1}{2}(\frac{c(1958)}{C})^2},$$

which reduces to:

$$P = \frac{1}{4\pi^2 B^2 C^2} e^{-\frac{1}{2}((\frac{b(1959)}{B})^2 + (\frac{b(1958)}{B})^2 + (\frac{c(1959)}{C})^2 + (\frac{c(1958)}{C})^2)}$$

The interesting part of this formula is the exponent, the part which depends on b and c. In order to maximize p with respect to k_1, k_2' and k_4, we try to bring the negative number in the exponent as close to zero as possible. This number is essentially the sum of the errors squared, exactly what we tried to maximize before. Once we have done this, it is well-known that we can maximize P by picking B and C to equal the root-mean-square average of b and c, respectively.

(vii) THE NEED FOR SOPHISTICATED NOISE MODELS

In general, there is little reason to believe that the classic normal distribution, of equations (2.4), will be a good model of the noise element in all social processes. Mosteller, for example, has pointed out that "flukes" occur fairly often in real social data.(23). There may be many processes which normally plod along in a predictable sort of way, governed by a noise process b(t) which fits a normal distribution and which never gets to be very large; every once in a while, however, the process may be hit by a fluke, which leads to changes much larger than one would have expected in the normal course of events. Suppose that "p_1" is the probability, at any time, of getting a fluke, Then the probability distribution for b(t) may actually fit this kind of equation:

$$p(b) = (1-p_1)\frac{1}{\sqrt{2\pi}B} e^{-\frac{1}{2}(\frac{b}{B})^2} + p_1 * \frac{1}{\sqrt{2\pi}B_1} e^{-\frac{1}{2}(\frac{b}{B_1})^2}$$

where B_1 is much larger than B. This equation states that most of the time - $(1-p_1)$ of the time, to be precise - b will fit the same bell-shaped curve as before; however, when a "fluke" occurs, b will fit a much broader bell-shaped curve, leading to much larger values for b. One way to account for these effects is to use this probability formula explicitly, instead of the usual normal distribution, in one's noise model; it may be impossible to

estimate "B_1" accurately, but, if flukes are a serious problem, it may still be possible to estimate k_1 and k_2 more accurately, and to show when "P" is larger for this kind of model.

Another source of noise, rarely handled explicitly in social science, is "measurement noise." In our discussion above, we talked about "A(1959)" and "A(1960)" as if we had available exact data for the true levels of assimilation in those years. We may have data, but there is good reason to believe that errors of different sorts occurred in the collection of this data. Even if the data were "perfect", in the sense of giving us a perfect measure of who speaks what language when, for example, they may still not be giving us a perfect measure of the underlying concept of assimilation, as governed by equations (2.3). Let us define "U(t)" as the true size of the unassimilated population, and "U'(t)" as the measured size of the unassimilated model. Then we might modify equation (2.3b) by writing:

$$\begin{aligned} U(t+1) &= k_3 U(t) + c(t) \\ U'(t+1) &= U(t+1) + d(t+1), \end{aligned} \qquad (2.16)$$

where "c" is the noise going on in the process itself, and where "d" is the measurement noise. The "process noise," "c", is a random factor in the actual process (top equation) which determines the objective evolution of the real variable we are interested in, U,

through time. The "measurement noise", "d", does not affect the objective reality, U, but only adds a factor of distortion to our measurement of U, our U';U'-U, the difference between the measured value of U and the true value of U, equals the measurement error, d. Even if U' did represent an objective variable in its own right, but a variable different from the one we really postulate to govern the dynamics, then this mathematical structure would still apply.

Given that we do not know the true value of U(t) at any times t, this model is not an "explicit" model; it does not tell us directly how to estimate U'(t+1) from earlier data available. (Note that the noise term, "c(t)", makes it impossible to calculate later values of U(t) from an estimate of U(0).) However, Box and Jenkins(24) have shown that this model is equivalent to the explicit model:

$$U'(t+1) = k_3U'(t) + f(t+1) - k_4f(t),$$

where "f" is a noise process fitting a normal distribution. This model has a kind of "memory term" in it, $k_4f(t)$, and may be estimated by use of the dynamic feedback method, as discussed in section (iv). In Chapter (III), we will describe how this method can be specialized to deal with models of this general sort, with any number of dependent variables. Economists, like Cochrane and Orcutt, have developed techniques to deal with some of the secondary consequences of measurement noise, like the problem of

serial correlation; however, in their original article(25), these authors have made it quite clear that the general problem of measurement noise is beyond the scope of their techniques.

This idea can be taken even further, if we wipe out the term for "process noise," and allow for the possibility of measurement noise only. In equation (2.16), this would mean eliminating the term c(t), while retaining d(t+1) and the other terms of the model; this would make (2.16) an "explicit" model, with memory U(t), similar to our example of section (iv). At first glance, this procedure sounds both unrealistic, and totally inferior to the procedures of the paragraph above. Process noise does exist in most social and ecological processes; as long as we can account for process noise and measurement noise both, why should we limit ourselves to the second possibility only?

Let us begin by seeing what this process really entails. If we assume that there is no process noise at all, then we can start out from our initial estimates (or data) for our variables, and solve our equations <u>exactly</u> to yield a stream of predictions for later data, right up to the end of our data set; these predictions account <u>only</u> for the data in the initial time-period. Notice that this is exactly what we were thinking of doing, early in section (ii), before we introduced the more "sophisticated" concept of ordinary regression. Also, note that this is what Jay Forrester's techniques(26) for "dynamic systems analysis"

tend to involve, even though he prefers to judge the final fit of his models by eye rather than by computer. Above all, note that the practical value of social and ecological models usually lies in their ability to predict the situation at distant times, without requiring knowledge of intervening times in the future. (This includes, of course, the ability to predict the results of different policies.) Using our new procedure, we evaluate models by their ability to yield good predictions across long periods of time, not by their ability to predict across the smallest possible period of time; therefore, we will generate models and coefficients better suited to the practical demands which will be placed on them. In order to estimate such models, we will have to resort to the dynamic feedback techniques of section (iv) above. The prior unavailability of such techniques offers at least some justification for the apparent disregard of empirical data in some of Forrester's more interesting work(27); however, if these estimation routines should become available soon at the Cambridge Project Consistent System, a more empirical analysis of these issues will become possible.

In short, the practical reasons for disregarding process noise can be very strong, at times. From a theoretical point of view, however, the reasons are not so obvious. If we start with a given statistical process, governed by a given set of equations, then those equations themselves are the best possible basis for prediction,

whether across one period of time or many. Thus "truth" implies predictive power. When we have a given set of data, the maximum likelihood method allows us to make use of all the information available - not just one measure, like long-term predictive power over the given data - to find the parameters and model closest to being "true."

The difficulty in this argument is that "closeness to truth," unlike "truth" itself, can be measured in many different ways. The Bayesian school of thought has begun to argue that this point, too, should be accounted for in practical statistical routines(28); however, their concepts of "loss function" do not fully encompass the concept of "long-term predictive power" here. As a practical matter, most models in the social sciences and in ecology are simplified, approximate models, which we do not expect to be "true" in any absolute sense; we only expect tham to approach truth, or, more realistically, we expect them to give us predictions similar in a broad way to what we would predict if we knew the full truth. Even when these models contain a hundred variables or so, they will still be hundreds of times simpler than the complete systems which they represent. If there were an infinite quantity of representative data available, and if we had to choose between a limited set of models, none of which are "true" in an absolute sense, then the model which performs best, on, say, predicting across ten years of time,

over this data, can also be expected to give us the best predictions of the future ten years hence. In order to estimate such a model, we should indeed try to minimize the errors across ten years, instead of following the conventional likelihood approach. In other words, if we wish to carry out an estimation which is "robust" - an estimation which will give us good predictions despite the oversimplifications of one's model - then a direct maximization of predictive power is appropriate; that is exactly what our "measurement noise only" approach entails.

In reality, we will have to accept limits both upon our choice of models and upon the size of our data. We will have two sorts of information to use in evaluating the predictive powers of our models: (i) the long-term predictive power as measured directly over the available data; (ii) general information about the "truth" of our model, as given by the maximum likelihood formulas. The first information is a direct measure of what we want to know. On the other hand, we only have a certain limited amount of this kind of information, in our data. The second information does tell us something about how close our model is to "truth", which in turn tells us something about predictive power. When our total information is limited, statistical theory recommends that we make use of all the information at our disposal, including both the "hard" and the "soft." Our problem, then, is one of a more practical nature:

which of the two sources of information should we emphasize, when we want to build a model suitable for medium-term and long-term prediction?

General guidelines for dealing with this problem will have to come from experience, experience with both ordinary procedures and with the new procedure suggested here. It should be clear, however, that the relative importance of process noise and measurement noise will vary from case to case. A direct comparison of the methods, say, in predicting the second half of one's data from the first half, would probably be desirable, in most cases. When the major flaw in one's existing model lies in its inability to describe measurement noise accurately, then one would suspect the possibility that the unexplained portion of that measurement noise would be organized enough to be partly "predictable" from one's process variables and noise; this would lead to distortion of the parameters of the process proper. For models which have this problem, the best way to improve predictive power may be to avoid this distortion, by making sure that measurement noise is not falsely attributed to the process equations (i.e. to process noise); by falling back on a "measurement noise only" model, in which process noise does not exist at all, one can eliminate this distortion entirely. Once again, as noted on the previous page, the "measurement noise only" approach involves no distortion at all,

insofar as it maximizes long-term predictive power, directly; its weakness lies in the lack of formal statistical efficiency. When process noise is very large, and the neglect of process noise would appear to seriously weaken one's ability to make full use of one's data, then it would still be possible to compromise, by the relaxation methods to be discussed in section (xi); these methods, by allowing process noise, but by making it much more "expensive" to attribute randomness to process noise than to measurement noise, may reduce the false attribution of the latter to the former, while preserving an adequate level of statistical efficiency. It is conceivable that in social science, as in hard science, there will someday be a viable distinction between "practical" statistical work, where prediction is most important, and "theoretical" statistical work, where "truth" as such turns out to be a more effective guide to finding new variables and terms to use in one's models. However, once again, the practical values of these techniques will have to emerge from empirical work. The empirical work of this thesis, in Chapters (IV) and (VI), does provide a strong indication that the "measurement noise only" approach is superior to the pure maximum likelihood approach, in the social sciences; this indication has been strong enough to totally reverse our own initial bias in favor of the classical approach. Still, the empirical studies here are only the beginning of a long process.

Finally, let us consider one other situation where the conventional approach to noise is inadequate: the case of "ideal types." Very often, in social science, we run across variables like "Republican President" and "Democratic President" which do not tend to vary across a continuous spectrum; they tend to be simply "true" or "false." The error in predicting such variables would not follow a normal distribution, but the problem need not be overwhelmingly difficult. On the other hand, we often find societies falling into certain distinct "ideal types"(29), such as "traditional", "developed" and "transitional"; we may find that a whole collection of other variables - political stability(30), aggressiveness, economic growth, etc. - depend heavily on which ideal type a society falls into. As an extreme example, let us imagine that there are three "ideal types" a nation might fall into, and that we have been studying four social variables which are all really determined by the current "ideal type":

	Type 1	Type 2	Type 3
Variable 1	1	0	1
Variable 2	0	1	1
Variable 3	1	0	0
Variable 4	0	1	0

Table II-3: Hypothetical Example of "Ideal Types".

If we predict that variable one will equal one, and discover that it actually equals zero, then this last piece of information tells us exactly what to expect for variables two, three and four. If we already made predictions for variables two, three and four, then we would also know now exactly what errors to expect in these predictions. Thus there is a connection, though complex, between the "errors" in predicting different variables. If our example had been somewhat more complex, with a lot of nonlinearity, and a certain amount of freedom to deviate from one's ideal type, it is clear that the correlations between the prediction errors of different variables could become hopelessly complex.

According to maximum likelihood theory, as sketched briefly in section (v), it is important to minimize the "right" measure of error, even when we estimate the coefficients we intend to use in making predictions of this process. The "right" measure is supposed to correspond to the actual noise process going on. If it does become important, in practice, that we do have such an accurate measure of error, and if the actual noise process is as complex as above, then we face serious difficulties in estimating any parameters at all. In our simple example, we could escape from these difficulties, by carrying out a simple factor analysis to detect the ideal types; then we could go on to study the ideal types only, and disregard

the original four variables. However, in the general case, a linear technique like factor analysis may not be enough; also, we may still want to consider the original variables, to account for whatever independent variation they have. In any case, it is clear that the conventional model of independent errors, following a normal distribution, cannot deal effectively with this kind of situation. The "measurement noise only" technique could conceivably reduce the difficulties here, but one would still expect a better model to emerge, if one could account for the complex interrelations of the process variables more explicitly.

In summary, in order to produce a "true" model of a social process, which is also capable of yielding good predictions, one must have an accurate model both of the "predictive part" (like equations (2.3)), and of the "noise part" (like equations (2.4)); otherwise, the standard techniques of statistical estimation may yield unrealistic estimates of both. If one pursues an unbalanced approach, giving more weight to the "predictive" part than to the "noise" part, one may soon find oneself in a situation where the inaccuracies in one's noise model are so large that any improvements in the "predictive" part are reflected by little improvement, if any, in the statistical likelihood of one's model. The "models without process noise" discussed above can, at the very least, serve as detectors for this kind of difficulty.

(viii) HOW TO ESTIMATE EXPLICIT SOPHISTICATED NOISE MODELS

Suppose that we had decided to make the Deutsch-Solow model for assimilation more sophisticated, not by working on the "predictive" part, but by working on the "noise" part; suppose that we decided to account for the possibility of "flukes," as discussed above. Then we might write the model:

$$A(t+1) = k_1 A(t) + k_2 U(t) + b(t+1)$$

$$p(b) = (1-p_1)\frac{1}{\sqrt{2\pi}\,B}\, e^{-\frac{1}{2}(\frac{b}{B})^2} + p_1 * \frac{1}{\sqrt{2\pi}\,B_1}\, e^{-\frac{1}{2}(\frac{b}{B_1})^2} \qquad (2.17)$$

Our problem is to try to maximize "P", which, as in the case of ordinary regression, will equal the product:

$$P = p(b(2))p(b(3))p(b(4))\ldots\ p(b(T)),$$

where T is the last time period for which we have data. An easier way to approach this is by trying to maximize the logarithm of P:

$$L = \log P = \log p(b(2)) + \log p(b(3)) + \ldots + \log p(b(T)).$$

We are trying to pick out the best possible values for the parameters k_1, k_2, B, B_1 and p_1. As before, we can try to do this by using "steepest descent"; as before, this means trying to measure the derivatives, $\frac{\partial L}{\partial k_1}$ etc. As before, in section (iii), we can set up a table of operations for each time t, which corresponds to a table which would emerge from a computer program to analyze this model;

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$\log p(b(t+1))$	32	logarithm	31	-
$p(b(t+1))$	31	sum	30	25
$\frac{(1-p_1)}{\sqrt{2\pi}\,B} e^{-\frac{1}{2}(\frac{b(t+1)}{B})^2}$	30	product	29	16
$e^{-\frac{1}{2}(\frac{b(t+1)}{B})^2}$	29	exponential	28	-
$-\frac{1}{2}(\frac{b(t+1)}{B})^2$	28	product	27	1
$(\frac{b(t+1)}{B})^2$	27	product	26	26
$\frac{b(t+1)}{B}$	26	ratio	20	9
$\frac{p_1}{\sqrt{2\pi}B_1} e^{-\frac{1}{2}(\frac{b(t+1)}{B_1})^2}$	25	product	24	13
$e^{-\frac{1}{2}(\frac{b(t+1)}{B_1})^2}$	24	exponential	23	-
$-\frac{1}{2}(\frac{b(t+1)}{B_1})^2$	23	product	22	1
$(\frac{b(t+1)}{B_1})^2$	22	product	21	21

Table II-4: Table of Operations for Equations (2.17)
(top section)

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$\frac{b(t+1)}{B_1}$	21	ratio	20	8
$b(t+1)=A(t+1)-k_1A(t)-k_2U(t)$	20	difference	6	19
$k_1A(t)+k_2U(t)$	19	sum	18	17
$k_1A(t)$	18	product	5	11
$k_2U(t)$	17	product	4	10
$\frac{1-p_1}{\sqrt{2\pi}\,B}$	16	ratio	15	14
$1-p_1$	15	difference	3	7
$\sqrt{2\pi}\,B$	14	product	2	9
$\frac{p_1}{\sqrt{2\pi}\,B_1}$	13	ratio	7	12
$\sqrt{2\pi}\,B_1$	12	product	8	2

Table II-4: Table of Operations for Equations (2.17)
(middle section)

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
k_1	11	parameter	-	-
k_2	10	parameter	-	-
B	9	parameter	-	-
B_1	8	parameter	-	-
P_1	7	parameter	-	-
A(t+1)	6	given	-	-
A(t)	5	given	-	-
U(t)	4	given	-	-
1	3	given	-	-
$\sqrt{2\pi}$	2	given	-	-
$-\frac{1}{2}$	1	given	-	-

Table II-4: Table of Operations for Equations (2.17) (bottom section)

this table is shown in Table II-4.

We may use such a table, as before, to calculate all the derivatives required by the steepest descent method. We may compute $\frac{\partial}{\partial k_1} \log p(b(t+1))$, etcetera, by inserting $S_{32}=1$, and working down the table as before to compute all the derivatives. The error model was complicated; therefore, the table is long. We can compute $\frac{\partial L}{\partial k_1}$ simply by computing $\frac{\partial}{\partial k_1} \log p(b(t+1))$ for all times t, from 1 to T-1, and adding up all the results; this would be the same sort of operation as in section (iii). In brief: if our model of error is complicated, but explicit, then dynamic feedback can be used to estimate the parameters of our model. Notice, if there had been <u>two</u> variables to predict, A(t+1) and U(t+1), that the <u>two</u> error terms b(t+1) and c(t+1) would appear somewhere in the middle of the table; if p(b(t+1),c(t+1)) were a function of <u>both</u> errors, a very complicated but explicit function, we could still have put together a table like this and used dynamic feedback. Also, if there were "memory" in the model, we could merge all these tables, for different times t, just as we did in section (iv). Note, however, that when the models become extremely complex, the choice of the initial guesses for our constants, to be used with the steepest descent method, becomes increasingly important; bit by bit, this problem becomes a subject worthy of attention in its own right, as our models grow in complexity.

(ix) PATTERN ANALYSIS

When ideal types or other systematic patterns are present, as in our example in section (vii), then it may be very difficult to formulate a good explicit model of noise, accounting for all the interrelationships between the errors in the predictions of different variables. A more natural way to handle such situations is by finding out what the ideal types are, and trying to predict them instead of predicting our original variables. In order to do this, we must try to find a way to describe the data at time t+1 in terms of a limited number of "ideal type" variables. Our description should be "complete", in the sense that we can regenerate the original data at time t+1 from knowing the ideal types, with minimum error. In the case of simple or only moderately complicated systems, with limited data available, we may use this approach as a way to reduce the number of variables, as we do with factor analysis. More generally, if we find that we have a large set of variables, heavily interconnected in a nonlinear way, we may try to find a set of "fundamental" variables which govern the behavior of all the original, more superficial variables, in a more independent, more linear and more comprehensible manner. The formulas which we use to estimate such variables might be considered to be "pattern detectors" or "feature detectors", in the language of pattern recognition. With a complex nonlinear system, the number of fundamental variables might actually be larger than the

original number of variables; however, it would still be much smaller than the number of possible system configurations.

Let us imagine that we start out with a set of variables to study, $X_1,\ldots,X_i,\ldots,X_n$, forming a vector, $\vec{X}$. We are looking for another set of variables, $R_1,\ldots,R_m$, forming a vector, $\vec{R}$, which "governs" the vector $\vec{X}$ in the sense that it accounts for all the cross-correlation between the different components of $\vec{X}$. (More precisely, it accounts for the cross-correlation in the random disturbances applied to the different components of $\vec{X}$.) At every time t+1, we wish to define these variables, $R_i(t+1)$, as functions of the data at time t+1, $\vec{X}(t+1)$. More generally, we may allow them to be functions of $\vec{X}(t)$ and $\vec{R}(t)$ also; this would allow us to detect dynamic patterns, involving such phenomena as population growth or physical motion.

Thus we may define:

$$R_i(t+1) = f_i(\vec{X}(t+1), \vec{X}(t), \vec{R}(t)).$$

In trying to "find" or to "define" the fundamental variables, R_i, our goal is to adjust the parameters of the functions f_i to fit the verbal requirements implied by our discussion above. These requirements involve the dynamic relations of the R_i and X_i variables; thus we can fit the parameters of the functions f_i only within the broader context of fitting a dynamic model of the entire process.

The first of the requirements we must meet is that the $R_i(t+1)$, unlike the original $X_i(t+1)$, are generated by independent stochastic processes. Our dynamic model must include a description of each of these processes. Thus it must specify the probability distribution for each variable, $R_i(t+1)$, as a function of $\vec{R}(t)$ and $\vec{X}(t)$; it must maintain the assumption that these probability distributions are independent of each other. Thus we may write:

$$p(R_i(t+1) \mid \vec{X}(t), \vec{R}(t)) = g_i(R_i(t+1), \vec{X}(t), \vec{R}(t)).$$

These functions, g_i, like the functions f_i, are part of our model. Rather than assume that we start out with the "correct" g_i, we will try to adjust the parameters of the functions g_i and the parameters of the functions f_i, both, in order to make the model as a whole fit the data as well as possible; this procedure will presumably adjust the functions f_i to fit as well as possible the assumption of independence, which is built into this model.

Finally, we have a second verbal requirement to meet. We require the ability to regenerate the $X_i(t+1)$ back again from $\vec{R}(t+1)$, with minimum possible error; as before, we can also allow the use of information from $\vec{R}(t)$ and $\vec{X}(t)$ in this procedure. In setting up equations to predict the $X_i(t+1)$, from known values of $\vec{R}(t+1)$, $\vec{X}(t)$ and $\vec{R}(t)$, we are effectively just extending our dynamic model to predict a new set of variables. We want the value of

$\vec{R}(t+1)$ to account for all the interdependence of the variables, $X_i(t+1)$; thus, once the value of $\vec{R}(t+1)$ is known, we want to be able to predict all of the $X_i(t+1)$ independently of each other. Thus we want to extend our dynamic model to describe the probability distribution of each variable, $X_i(t+1)$, as a function of $\vec{R}(t+1)$, $\vec{R}(t)$ and $\vec{X}(t)$; we want to maintain the requirement that each of these probability distributions is independent of all the others. Thus we may write:

$$p(X_i(t+1) \mid \vec{R}(t+1), \vec{R}(t), \vec{X}(t)) = h_i(X_i(t+1), \vec{R}(t+1), \vec{R}(t), \vec{X}(t))$$

These functions, h_i, like the functions g_i and f_i, are part of our model. In adjusting the parameters of all these functions, to fit the data, we will hope to adjust the parameters of the f_i to fit the assumptions of independence both for the g_i and for the h_i.

Our objective, then, is to estimate the functions f_i, g_i and h_i, so as to maximize the likelihood of this model as a whole. In order to do this, we could calculate the likelihood as we have with other models:

$$p(\vec{X}(t+1) \mid \vec{X}(t), \vec{R}(t), \text{model}) = p(\vec{X}(t+1) \mid \vec{R}(t+1), \vec{X}(t), \vec{R}(t), \text{model}) * p(\vec{R}(t+1) \mid \vec{X}(t), \vec{R}(t), \text{model})$$

(Notice, in this equation, that we do not have to integrate over all possible values of $\vec{R}(t+1)$, on the right, because the $R_i(t+1)$ have been defined as definite functions of the other variables here; it is as if they were components of $\vec{X}(t+1)$, or, from another point of

view, as if their probability distribution contingent on $\vec{X}(t+1)$, $\vec{R}(t)$ and $\vec{X}(t)$ were a Dirac delta function which we have already integrated implicitly.) This yields a likelihood measure for a complete set of data:

$$
\begin{aligned}
L &= \sum_t \log p(\vec{X}(t+1) \mid \vec{X}(t), \vec{R}(t)) \\
&= \sum_t \Big(\sum_{i=1}^{n} \log h_i(X_i(t+1), \vec{R}(t+1), \vec{X}(t), \vec{R}(t)) \\
&\qquad + \sum_{i=1}^{m} \log g_i(R_i(t+1), \vec{X}(t), \vec{R}(t)) \Big) \\
&= \sum_t \Big(\sum_{i=1}^{n} \log h_i(X_i(t+1), \vec{f}(\vec{X}(t), \vec{R}(t)), \vec{X}(t), \vec{R}(t)) \\
&\qquad + \sum_{i=1}^{m} \log g_i(f_i(\vec{X}(t), \vec{R}(t)), \vec{X}(t), \vec{R}(t)) \Big)
\end{aligned}
$$

Using this likelihood function, we may construct tables, analogous to those used in section (iii) for log p, to let us calculate the derivatives of likelihood with respect to all of our parameters, in the functions f_i, g_i and h_i. Thus, once again, we may use the method of steepest descent to maximize likelihood.

It should be emphasized, however, that the likelihood function spelled out in the equation above was based on substitutions which were, in some ways, arbitrary. From a formal point of view, the functions f_i, g_i and h_i are somewhat redundant as model specifications; thus we have a certain amount of leeway in deciding how to combine them. When, as above, the functions h_i are adjusted

in such a way that the "f_i" are considered to be fixed but effectively unknown functions, and in such a way that $\vec{R}(t)$ and $\vec{X}(t)$ are directly available to the h_i as arguments, then the resulting model will, as a whole, be at least as good as a simple model specifying the $X_i(t+1)$ as independent functions of $\vec{R}(t)$ and $\vec{X}(t)$; in other words, even if the $R_i(t+1)$ are totally ignored, a model fit in this way can achieve, at a minimum, the level of fit that would be achieved by a conventional model assuming independence. In order to get maximum value from this technique, however, one would want to adjust the definition of the "f_i" to increase the actual likelihood of the model as a whole, evaluated in terms of the observed data, $\vec{X}(t+1)$, by themselves. It is likely that the constraint of having to assume independence at all levels, in order to minimize cost with a large number of variables, might not be consistent with achieving an absolute maximum of likelihood by this more strenuous criterion. Also, it is far from obvious that the procedure above is the best procedure for measuring likelihood, even subject to that constraint. The concepts of time-series analysis discussed elsewhere in this thesis required considerable theoretical and empirical work, both, before the pros and cons of specific algorithms began to seem clear; pattern analysis, which is a more subtle and potentially more powerful technique, will require at least as much development, both theoretical and empirical, to become useful in the future. Theoretical

studies of the linear special case may be of particular value in the early stages of this development.

Even at this stage of research, however, it seems clear where the applications of pattern analysis will lie. Pattern analysis is essentially a generalization of the idea of factor analysis to the nonlinear dynamic case. The dynamic power of a proposed "principal factor" would appear to be a better measure of its importance than the variance it account for in static cross-sections; when time-series data are available, pattern analysis would appear to be a clearly superior strategy for evaluating the same set of parameters as with factor analysis. With many variables, or long time-series, the nonlinear feature may also turn out to grow in importance; statistical pattern-recognition, or satellite-collected data, may both provide major applications for the possibility of nonlinearity here. In such highly complex systems, the massive number of variables may make the assumption of independence a necessity, both in terms of computational cost and in terms of avoiding models with more degrees of freedom than one could hope to estimate; pattern analysis may be essential to prevent excessive reductions of model likelihood as a result of that assumption. Also, with such systems, note that one does not have to restrict one's computer package to estimating functions - f_i, g_i and h_i - whose form has been specified in advance by the user. One can provide an option for the computer to

try out new tables of operations, automatically, by pruning out terms which contribute little and by evaluating the improvement in likelihood from adding new terms, chosen essentially at random. Finally, one should note that the assumption of independence may be especially valuable on machines which allow parallel processing; the one-to-one association between functions f_i and functions g_i, corresponding to the same components of $\vec{R}$, may be of major importance in making pattern analysis operational on such machines.

(x) OPTIMIZATION

In sections (iii) and (iv), we saw how dynamic feedback can be used to minimize error; later on, we saw how it could be used to maximize probability. In general, the method of steepest descent can be used to minimize or maximize any function we please, so long as we can calculate all the derivatives. Dynamic feedback lets us calculate the derivatives, so long as our system of formulas is explicit. Therefore: the dynamic feedback method can be used to minimize or maximize other things besides error.

Suppose, for example, that we have a simple model of the US economy, something like this:

$$\begin{aligned} C(t+1) &= k_1 C(t) + k_2 Y(t+1) \\ Y(t+1) &= a_1 P(t+1) \\ P(t+1) &= P(t) + k_3(P(t) - C(t)) \end{aligned} \qquad (2.18)$$

In this case, we are not trying to evaluate or estimate a model. We assume that the model has already been tested, and that k_1, k_2 and k_3 have already been estimated by some kind of statistical procedure. "C" here represents consumption; "Y" represents personal income; "P" represents production. With optimal government policy, all production capacity will be channelled to either consumption or to some kind of investment; "a_1", the rate of taxation, determines how much goes to each. Our problem here is to find the "best" level for "a_1". Suppose that we define "best" to mean the level of a_1 which maximizes consumption in the long term. Suppose that we start from a known position in year 1, and want to maximize the sum of consumption over the next three years. Then we may define our utility function, "U", to equal C(2)+C(3)+C(4). We may set up the table of calculations, shown in Table II-5 over the next few pages, which defines how "U" is to be calculated up from the parameter and the constants of this problem. In order to maximize U by the method of steepest ascent, we need only calculate $\frac{\partial U}{\partial a_1}$, the derivative of U with respect to the parameter we have to control. We may calculate the derivatives of $U=X_{29}$, as before, by using the method of dynamic feedback on the table of operations, Table II-5. We may start with $S_{29}=\frac{\partial U}{\partial X_{29}}=1$; then, if we calculate derivatives down the table, as before, S_1 will equal $\frac{\partial U}{\partial a_1}$. Our original model was very simple in this example; however, it should be clear that

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$U = C(2)+C(3)+C(4)$	29	sum	28	21,14
$C(4) = k_1C(3)+k_2Y(4)$	28	sum	27	26
$k_1C(3)$	27	product	21	4
$k_2Y(4)$	26	product	25	3
$Y(4)=a_1P(4)$	25	product	24	1
$P(4)=P(3)+k_3(P(3)-C(3))$	24	sum	23	17
$k_3(P(3)-C(3))$	23	product	22	2
$P(3)-C(3)$	22	difference	17	21
$C(3) = k_1C(2)+k_2Y(3)$	21	sum	20	19
$k_1C(2)$	20	product	14	4
$k_2Y(3)$	19	product	18	3
$Y(3)=a_1P(3)$	18	product	17	1
$P(3)=P(2)+k_3(P(2)-C(2))$	17	sum	16	10
$k_3(P(2)-C(2))$	16	product	15	2
$P(2)-C(2)$	15	difference	10	14

Table II-5: Table of Operations for Equations (2.18). (top section)

Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$C(2) = k_1 C(1)+k_2 Y(2)$	14	sum	13	12
$k_1 C(1)$	13	product	7	4
$k_2 Y(2)$	12	product	11	3
$Y(2)=a_1 P(2)$	11	product	10	1
$P(2)=P(1)+k_3(P(1)-C(1))$	10	sum	9	5
$k_3(P(1)-C(1))$	9	product	8	2
$P(1)-C(1)$	8	difference	5	7
$C(1)$	7	given	-	-
$Y(1)$	6	given	-	-
$P(1)$	5	given	-	-
k_1	4	given	-	-
k_2	3	given	-	-
k_3	2	given	-	-
a_1	1	parameter	-	-

Table II-5: Table of Operations for Equations (2.18). (bottom section)

even a complicated nonlinear model, involving many control parameters, could be translated into a table like Table II-5, by computer, if the model is "explicit" in the sense of section (iv).

In our example above, we have described a problem which does not quite fit the standard format used most often in control theory. It is a problem in what we would prefer to call "systems optimization" or "dynamic systems optimization." Problems of this type have been discussed by Jacobson and Mayne(31), as a device for overcoming some of the difficulties of optimization under conditions of uncertainty; in the social sciences, however, this formulation may have substantive advantages over the more standard formulation, which are worth pointing out here.

In the case above, for example, we tried to pick the best possible value for "a_1", a <u>constant</u> in the nation's taxation system. In standard control theory, one would usually look at $a_1(t)$, the taxation rate <u>at each time</u>, and try to find the best possible schedule of tax rates for different years. In principle, the second way is better, but only if it is feasible, politically, to change the tax rates up and down every year. In practice, governments trying to follow conventional Keynesian policies, adjusting tax rates every year, have encountered serious political problems and problems of timing; thus there has been great interest in "automatic adjustment" factors(31), and in other system parameters which can be adjusted to

improve economic performance without forcing us to change policy too often. Thus "systems optimization" has something worthwhile to offer the policy-maker, above and beyond its mathematical convenience.

The methods here could also be used in conventional control theory problems. In our example, we could try to pick the best values for the three parameters, a(2), a(3) and a(4), by putting all three at the bottom of our table, and calculating back $\frac{\partial U}{\partial a(2)}$, $\frac{\partial U}{\partial a(3)}$ and $\frac{\partial U}{\partial a(4)}$. Jacobson and Mayne(32) have shown how steepest ascent methods, very similar to ours, can also be used in cases where noise terms appear in the model. The dynamic feedback method allows us to calculate $\frac{\partial U}{\partial a_i}$, in cases where the dynamic laws of the system are arbitrarily complex, and where the interconnections in time may stretch over several time-periods; it allows us to exploit the internal structure of the model, as spelled out in a table, in order to calculate back all the derivatives in one pass, at a cost much lower than with separate differentiations. Otherwise, however, the methods discussed by Jacobson and Mayne for making use of $\frac{\partial U}{\partial a_i}$, in systems optimization, are very general, and do not require further elaboration here.

(xi) THE METHOD OF "RELAXATION" WITH MEASUREMENT-NOISE-ONLY MODELS

In section (vii), we have discussed the possibility of "measurement-noise-only" models, which are at the opposite pole from the usual regression models, which may be characterized as "process-noise-only" models. Between these two poles is a whole spectrum of more moderate techniques. Let us suppose that one has a simple model of some process, defined by the equations:

$$X_i(t+1) = f_i(X_1(t),\ldots X_n(t)) \qquad i = 1, n \qquad (2.19)$$

Using the classical regression approach, we would tack on a normal noise term, $a_i(t)$, to the end of these equations, to arrive at the stochastic model:

$$X_i(t+1) = f_i(X_1(t),\ldots X_n(t)) + a_i(t)$$

We would estimate the parameters in this model by trying to minimize the square errors:

$$L_i = \sum_t (X_i(t+1) - f_i(X_1(t),\ldots X_n(t)))^2, \qquad (2.20)$$

after we substitute in for the <u>measured</u> values of $X_i(t+1)$ and $X_j(t)$. In effect, this would imply minimizing the square error for predictions over <u>one interval</u> of time.

With the measurement-noise-only models, we would normally include, in the list of parameters to be estimated, the values of $Y_i(0)$, where Y_i is defined as the "true" value of X_i. Using

equations (2.19), we can predict the "true" value of $X_i(1)$, $Y_i(1)$, from our estimates of the $Y_i(0)$, and predict $Y_i(2)$ from the predictions of $Y_i(1)$, et cetera. Using these long-term predictions, $\hat{Y}_i(t)$, we can try to minimize the errors:

$$L_i' = \sum_t (X_i(t+1) - f_i(\hat{Y}_1(t),\ldots \hat{Y}_n(t)))^2 \quad (2.21)$$

From the viewpoint of maximum likelihood theory, these predictions, $\hat{Y}_i(t)$, may be viewed as the "estimates" of $Y_i(t)$, derived from the estimates $\hat{Y}_i(0)$ and from the assumption that equations (2.19) are <u>exactly</u> true, with no noise, for the true values, $Y_i(t)$.

How can we find a viable compromise between (2.20) and (2.21)? In (2.20), we use the <u>measured</u> value, $X_i(t)$, to estimate the true values, $Y_i(t)$, for use as the arguments of f_i; in (2.21), we use estimates, $\hat{Y}_i(t)$, based solely on updating our estimates for $Y_i(t-1)$, i.e. $\hat{Y}_i(t-1)$, by use of (2.19). The obvious compromise is to estimate $Y_i(t)$ by something half-way between the measured values, $X_i(t)$, and the estimates of $Y_i(t)$ which result from updating our estimates of $Y_i(t-1)$. Thus we may define new estimates, $Z_i(t)$, of $Y_i(t)$, by:

$$Z_i(t) = (1-r)f_i(Z_1(t-1),\ldots Z_n(t-1)) + rX_i(t). \quad (2.22)$$

Using these estimates, we may attempt to minimize the loss function:

$$L_i'' = \sum_t (X_i(t+1) - f_i(Z_1(t),\ldots Z_n(t)))^2 \quad (2.23)$$

The $Z_j(0)$, like $Y_j(0)$, would be parameters to estimate. The constant "r" may be called the "coefficient of relaxation"; it is a kind of interest rate which, when large, implies a greater concern for short-term prediction than for long-term prediction.

Note that the structure of equation (2.22) looks similar to that of a filtering system, designed to yield a posterior estimate of the "true" value of $\vec{X}$, given both a prior expectation and an actual measurement. In the field of engineering, a great deal of work has been done on the problem of designing an optimal filtering system, to deal with vectors, $\vec{X}$, which result from noisy measurements of a linear process which has been completely specified in advance. It is well-known that the best way to update one's estimates, in this situation, is not by the independent equations (2.22), but by the matrix equations of the "Kalman filter"(33):

$$\vec{Z}(t) = \vec{f}(\vec{Z}(t-1)) + K(t)(\vec{X}(t)-H(t)\vec{f}(\vec{Z}(t-1))),$$

where K and H are time-varying matrices determined by:

$$K(t) = P(t)H^T(t)R^{-1}(t)$$

$$P(t) = (M^{-1}(t) + H^T(t)R^{-1}(t)H(t))^{-1}$$

$$M(t+1) = \phi(t)P(t)\phi^T(t) + G(t)Q(t)G^T(t),$$

and where H, R, G and Q are all characteristic matrices of the linear process, a process which may be specified:

$$\begin{aligned} \vec{X}(t) &= H(t)\vec{Y}(t) + \vec{b}(t) \\ \vec{Y}(t+1) &= G(t)\vec{Y}(t) + \vec{c}(t), \end{aligned} \qquad (2.24)$$

with R and Q the covariance matrices of the noise vectors $\vec{b}$ and $\vec{c}$, respectively. The details of these equations are beyond the range of our discussion here. One should note that the linear processes of equation (2.24) are essentially the same as those we will discuss early in Chapter (III); it will be shown in that chapter that processes of that general sort can be dealt with exactly by use of the "ARMA" approach, whose practical limitations will be discussed in Chapters (IV), (V) and (VI). However, even if the Kalman filtering equations were derived for a limited class of linear processes, one might expect them to be an improvement over equations (2.22), on the theory that they can be used to perform the same function, somewhat more rationally, as part of our system of robust estimation. In this case, one would adjust the matrices H, R and Q in an ad hoc sort of way, just as one would adjust the relaxation constant, "r", rather than estimate them all beforehand by use of the maximum likelihood technique on some version of the simple equations (2.24). However, this use of the Kalman filter brings three difficulties with it, which make it a subject for future research rather than present systems design: (i) the need to adjust three matrices, H, R and Q, automatically, requires a much greater development of the theory of robust estimation than does the need to adjust a single constant, r, by hand; (ii) the sheer complexity of the Kalman equations would impose

heavy costs on the systems programmer; (iii) even given a rational approach to estimating the matrices H, R and Q, one would presumably need a huge quantity of data to estimate so many parameters, in addition to all the parameters of one's model.

(xii) THE ORDERED DERIVATIVE AND DYNAMIC FEEDBACK

The traditional formalism used for dealing with partial derivatives was evolved to deal with the problems of geometry and of physical science. In those fields, one normally deals with functions defined over a fixed set of coordinate variables; even when one changes one's choice of coordinates, one is usually making a clearcut shift from one set to a second set. In the social sciences, however, one normally deals with a complex web of functional relations and variables. This web will often have a causal ordering associated with it. Thus when we say that $x_{t+1}=f(y_t,z_t)$, we not only mean that a relation exists between the variables y_t, z_t and x_{t+1}; we also tend to mean that the variables y_t and z_t "cause" x_{t+1} to equal what it does, and that x_{t+1} is causally "later" than y_t and z_t. We will often be interested in asking what changes will follow, later, if we change a given variable by a small amount at a given time. Clearly, this question calls for us to calculate some kind of partial derivative. In order to deal with

this kind of situation, as easily as we now deal with situations in physical science and geometry, we need to define a new formalism for this kind of partial derivative.

Let us begin by imagining that we have a well-ordered set of variables, $x_1, x_2, \ldots x_n$, with each variable x_i obeying a functional relation:

$$x_i = f_i(x_{i-1}, x_{i-2}, \ldots x_1).$$

Let us define a new set of functions, F_i, recursively:

$$\text{(i)}\quad F_n(x_n, x_{n-1}, x_{n-2}, \ldots x_1) = x_n$$

$$\text{(ii)}\quad F_{i-1}(x_{i-1}, x_{i-2}, \ldots x_1)$$

$$= F_i(\, f_i(x_{i-1}, x_{i-2}, \ldots x_1),\, x_{i-1}, x_{i-2}, \ldots x_1)$$

(In other words, "F_i" expresses x_n as a function of the variables $x_i, x_{i-1}, \ldots x_1$, arrived at by substitution into higher F_j's.)
Let us define the ordered derivative of x_n as follows:

$$\frac{\partial^+ x_n}{\partial x_i} = \frac{\partial F_i}{\partial x_i}, \qquad n \geq i \geq i_0$$

where the derivative on the right is evaluated by traditional procedures, holding constant all the variables $x_{i-1}, \ldots x_1$.

We may further define:

$$\frac{\partial^{+} x_n}{\partial x_i} = \frac{\partial F_{i_0}}{\partial x_i} \qquad 1 \leq i \leq i_0$$

Theorem:

$$\frac{\partial F_j}{\partial x_i} = \sum_{k=j+1}^{n} \frac{\partial^{+} x_n}{\partial x_k} \cdot \frac{\partial f_k}{\partial x_i} \qquad \text{for} \quad \begin{matrix} i_0 \leq j < n \\ 1 \leq i \leq j \end{matrix}$$

We can prove this, for any given i and n within the acceptable range, by induction on j downwards from j=n-1 (down to j=i). Let us begin by considering the initial case, j=n-1. In this case, our general claim reduces to:

$$\frac{\partial F_{n-1}}{\partial x_i} = \frac{\partial^{+} x_n}{\partial x_n} \cdot \frac{\partial f_n}{\partial x_i} \ .$$

From our definitions of F_n and of F_{n-1}, this reduces immediately to:

$$\frac{\partial f_n}{\partial x_i} = \frac{\partial x_n}{\partial x_n} \cdot \frac{\partial f_n}{\partial x_i} \ ,$$

which is clearly true.

Now, to complete our proof, we need only prove the formula for $j \geq i$ and $\geq i_0$, on the assumption that it is true for $j+1 \leq n$. Let us begin by going back to the definition of F_j:

$$F_j(x_j, x_{j-1}, \ldots x_1) = F_{j+1}(f_{j+1}(x_j, x_{j-1}, \ldots x_1), x_j, \ldots x_1)$$

In order to make this more explicit, we may write it as follows:

$$F_j(x_j, x_{j-1}, \ldots x_1) = F_{j+1}(s_{j+1}, s_j, s_{j-1}, \ldots s_1)$$

$$\text{where} \quad s_{j+1} = f_{j+1}(x_j, x_{j-1}, \ldots x_1)$$

$$s_i = x_i \qquad 1 \leq i \leq j.$$

By the conventional chain rule for partial differentiation:

$$\frac{\partial}{\partial x_i}(F_j(x_j, x_{j-1}, \ldots x_1)) = \sum_{k=1}^{j+1} \left(\frac{\partial}{\partial s_k}(F_{j+1}(s_{j+1}, \ldots s_1))\right)\left(\frac{\partial s_k}{\partial x_i}\right)$$

$$1 \leq i \leq j$$

Now for $k \leq j$, our definitions of s_k as a simple function of the x_i clearly tell us that $\frac{\partial s_k}{\partial x_i}$ equals 1 if k=i, and zero otherwise. For k=j+1, our definition tells us that: $\frac{\partial s_k}{\partial x_i} = \frac{\partial f_{j+1}}{\partial x_i}$. Thus the sum on the right in the equation above may be evaluated to give us a new equation:

$$\frac{\partial F_j}{\partial x_i} = \left(\frac{\partial}{\partial s_{j+1}}(F_{j+1}(s_{j+1}, \ldots s_1))\right)\frac{\partial f_{j+1}}{\partial x_i}$$

$$+ \frac{\partial}{\partial s_i}(F_{j+1}(s_{j+1}, \ldots s_1))$$

Given that our remaining derivatives with respect to the s_i do not involve the expression "x_i" anywhere, and given that the s_i have been defined in such a way as to equal the x_j for $1 \leq i \leq j+1$, the value of these expressions will not change if we substitute in the letter "x"

for every occurrence of the letter "s". Thus we get:

$$\frac{\partial F_j}{\partial x_i} = \frac{\partial F_{j+1}}{\partial x_{j+1}} \cdot \frac{\partial f_{j+1}}{\partial x_i} + \frac{\partial F_{j+1}}{\partial s_i}$$

Now from the induction hypothesis we were given that:

$$\frac{\partial F_{j+1}}{\partial s_i} = \sum_{k=j+2}^{n} \frac{\partial^+ x_n}{\partial x_k} \cdot \frac{\partial f_k}{\partial x_i}$$

From our definition of the ordered derivative, this is merely:

$$\frac{\partial F_j}{\partial x_i} = \frac{\partial^+ x_n}{\partial x_{j+1}} \cdot \frac{\partial f_{j+1}}{\partial x_i} + \sum_{k=j+2}^{n} \frac{\partial^+ x_n}{\partial x_n} \cdot \frac{\partial f_k}{\partial x_i}$$

$$= \sum_{k=j+1}^{n} \frac{\partial^+ x_n}{\partial x_n} \cdot \frac{\partial f_k}{\partial x_i},$$

which establishes our contention for the case j, and which, by induction, proves the contention as a whole.

Corollary 1: If $i_0 \leq i < n$:

$$\frac{\partial^+ x_n}{\partial x_i} = \sum_{k=i+1}^{n} \frac{\partial^+ x_n}{\partial x_k} \cdot \frac{\partial f_k}{\partial x_i}$$

This follows immediately from setting j=i in our theorem, and exploiting the definition of the ordered derivative.

Corollary 2: If $1 \leq i \leq i_0$:

$$\frac{\partial^+ x_n}{\partial x_i} = \frac{\partial F_{i_0}}{\partial x_i} = \sum_{k=i_0+1}^{n} \frac{\partial^+ x_n}{\partial x_k} \cdot \frac{\partial f_k}{\partial x_i} .$$

Notice that $F_{i_0}(x_{i_0},\ldots x_1)$ is the function which expresses x_n as a function solely of the "external parameters," x_{i_0} through x_1. If x_n represents something like "likelihood", L, and if the "external parameters" represent constants of the model and fixed data, then F_{i_0} expresses likelihood as a function of these parameters. When we are trying to maximize likelihood "as a function of these parameters," we are trying to maximize L <u>expressed as</u> F_{i_0}. Thus, when we ask about "$\frac{\partial L}{\partial k_1}$", in that context, we are really asking about $\frac{\partial F_{i_0}}{\partial k_1}$.

Notice that the concept of "ordered derivative" does not really depend upon the exact choice of order $x_1, x_2, \ldots x_n$. Suppose that "x_i" is really "simultaneous" with $x_{i+1}, x_{i+2}, \ldots x_{k-1}$, in the sense that it is not really an argument of the functions $f_{i+1}, f_{i+2}, \ldots f_{k-1}$. Then, in our chain rule above, the derivative $\frac{\partial f_j}{\partial x_i}$ is zero for j of i+1 through k-1; thus the actual value of the ordered derivative, as given by those formulas, will not be affected by our arbitrary decision to treat these variables as if they were "later" than x_i in the causal ordering. The ordered derivative would appear to be defined with respect to the general causal ordering, a weak ordering of our lattice of variables, rather than the strong numerical order chosen to represent it. For our purposes, however, it is not necessary to establish such generality, since we need only justify a calculating procedure based upon the definite numerical ordering chosen for our tables.

As a practical matter, all of the "working back" of derivatives cited above might be carried out by one standard computer subroutine, called on by simple "main programs" within one's computer package to carry out estimation and optimization for models of all different sorts. Other possibilities have been mentioned briefly in section (iii). Model specification could be allowed either in terms of standard TSP formulas, or in terms of Forrester-style "DYNAMO" expressions. The standard subroutine(s), set up to allow optimal control calculations, would, in principle, also allow maximum-likelihood estimation of "hidden variables" in <u>implicit models</u>; however, the use of this provision should probably not be encouraged, except in those cases where a theoretical understanding exists of its potential value.

APPENDIX: VARIATIONS ON THE STEEPEST ASCENT METHOD FOR EFFICIENT CONVERGENCE

The discussion of the dynamic feedback method throughout this chapter depends on the assumption that the derivatives calculated by this method can be used as the input to the steepest ascent method, in minimizing or maximizing various types of functions. In practice, however, we have found great difficulties in getting adequate convergence with the classical steepest ascent method, in our early experiments with ARMA estimation, to be discussed in Chapter (III). This experience seems to be in line with the general impressions of other people in the community who have used the method. It is possible, however, to bring the convergence rate up to reasonable standards by use of "variable metric techniques" and related methods.

In section (iii), we alluded briefly to the constants "C" to be used in the steepest ascent method; in our discussion, there was never any reason to require that "C" be the same for all of the parameters, a_i. In the "variable metric" approach, one simply chooses different constants, C_i, for different parameters. Thus one would write:

$$a_i^{(n+1)} = a_i^{(n)} + C_i \frac{\partial U}{\partial a_i}(\vec{a}^{(n)}).$$

This equation specifies that our n-plus-first estimate of the parameter a_i will equal our nth estimate, plus C_i times the

derivative with respect to a_i of the function, U, which we are trying to maximize. (The derivative is calculated, of course, from our current set of estimates, $\vec{a}^{(n)}$.) This approach has become fairly popular , in some quarters.

Ideally, if one had all the second derivatives available, one might use the classic Gaussian method:

$$\vec{a}^{(n+1)} = \vec{a}^{(n)} - A^{-1}(\vec{\nabla U}),$$

where "$\vec{\nabla U}$" is the vector $\frac{\partial U}{\partial a_i}$ and "A" is the matrix of the second derivatives, $\frac{\partial^2 U}{\partial a_i \partial a_j}$. To pick the constants, "C_i," above, one might try to pick them to form as close an approximation as possible to the Gaussian equation here. Thus one might try to approximate:

$$C_i = -1/(\frac{\partial^2 U}{\partial a_i^2}) .$$

In order to generate a low-cost, order-of-magnitude estimate, S_i^*, of $(\frac{\partial^2 U}{\partial a_i^2}) = (\frac{\partial^2 U}{\partial x_i^2})$, one might carry out another feedback calculation down our tables of operations:

$$S_i^* = \sum_{j=i+1}^{n} \left((\frac{\partial f_j}{\partial x_i})^2 S_j^* + (\frac{\partial^2 f_j}{\partial x_i^2}) S_j \right), \qquad (2.25)$$

where S_j is the ordered derivative of U with respect to x_j, as computed by the procedures of section (iii), and the rest of the notation here comes from section (xii). The term on the left side of the expression to be summed preserves the sign of the "estimated"

second derivatives, as we go down from the S^*_j to S^*_i. The term on the right, however, risks a change of sign; it might either be eliminated, or else cut off to equal zero whenever it tries to go too far in an abnormal direction, if we choose to avoid this risk. Note, in the case of a maximization problem, that the normal sign for the second derivatives is negative.

If we write:

$$C_i = -w/S^*_i,$$

then, if the n-plus-first estimate turns out to be inferior to the nth, we can simply reduce the unsubscripted constant "w" and try again. This method is guaranteed eventual convergence to a local maximum, as we adjust "w" back and forth, for exactly the same reasons that the classical method is guaranteed convergence as "C" is adjusted back and forth; given that we have imposed measures to keep the signs of the S^*_i negative, we may invoke the definition of the derivative, just as we did in section (iii).

In the case of ARMA estimation, a similar though more specialized approach has turned out to be quite successful. While we have not run across this particular form of the variable metric approach in the literature, we have heard rumors that something similar may have attracted attention elsewhere in statistics; however, it would be difficult to believe that equation (2.25), itself, which is based on

a procedure related to dynamic feedback, has been used in this general form.

In situations where the number of variables is great, and many iterations are required in any case, one can imagine an additional provision, "convergence learning," to try to make the constants C_i better approximations to the choice which would lead to the fastest convergence. One may set:

$$C_i = -w\theta_i/S_i^*,$$

where "θ_i" is increased when the parameter $a_i^{(n)}$ seems to be moving systematically in one direction from estimate (n) to estimate, and where it is decreased when $a_i^{(n)}$ seems to be oscillating. One might, for example, multiply θ_i by 1+C, for some positive C, when $\frac{\partial U}{\partial a_i}$ calculated at $\vec{a}^{(n+1)}$ has the same sign as $\frac{\partial U}{\partial a_i}$ calculated at $\vec{a}^{(n)}$; one might divide it by 1+C when the signs are opposite. As before, if the sign of θ_i is positive, one is still assured of eventual convergence to a local maximum. Insofar as each of these factors, "θ_k", is essentially an adjustment factor for our approximation of $(\frac{\partial^2 U}{\partial a_k^2})$, we might even use it in (2.25), to divide those terms in our summation on the right which involve entries, X_j, in our table of operations, which use the parameter a_k as a "source variable."

In order to define these procedures in more detail, it would be

necessary to have some way to evaluate the many alternative possibilities in these directions; insofar as these procedures are all aimed at the practical goal of speeding up convergence, it would seem best to evaluate them by way of practical experiments, when the necessary computer routines become available.

FOOTNOTES TO CHAPTER (II)

(1) Deutsch, Karl W., Nationalism and Social Communications, MIT Press, Cambridge, Mass. 1966, revised second edition, Appendix V. Note that we have used the letter "U" instead of "D", in the revised version of the model. Also note that several versions of this model have appeared in print. The version here, in all fairness, was actually taken directly from Hopkins, Raymond and Carol, "A Difference Equation Model for Mobilization and Assimilation Processes", 1969, unpublished; a copy of this paper was provided to us by Prof. Deutsch, and described as containing the final revision of the model. This revision appears, in difference equation form, in Hopkins, Raymond, "Projections of Population Change by Mobilization and Assimilation", Behavioral Science, 1972, p.254. The reasons for the revisions to earlier versions are described in Hopkins, Raymond, "Mathematical Modelling of Mobilization and Assimilation Processes", in Mathematical Approaches to Politics, edited by Hayward Alker, Karl Deutsch and Antoine Stoetzel, Elsevier Publishing Co., NY, 1973, p. 381.

(2) Rapoport, Anatol, Fights, Games and Debates, U. of Michigan Press, Ann Arbor, Mich., 1961, Second Printing, p. 173. The "prisoner's dilemna", in its original form, is a simple two-person game in which each player has two options to choose from: to betray or not to betray the other player to the police. If neither player is betrayed, both pay a slight penalty (a small jail sentence). If one is betrayed, then he pays a heavy penalty, but the other escapes all penalty. If both betray each other, both pay a fairly heavy penalty. The structure of this game has been used as a paradigm for certain arms races, in which the self-interest of each player, paradoxically, may lead both into a competition in which both of them enjoy less security and have less money left over than if both had shown restraint.

(3) Even the most elementary models used in economics tend to be used to generate tangible numerical prediction; see, for example, Economics : An Introductory Analysis, by Paul A. Samuelson, Fourth Edition, McGraw-Hill, NY, 1958, chapters eleven through thirteen. More explicit predictive models may be found in Hickman, Bert G., Econometric Models of Cyclical Behavior, National Bureau of Economic Research, 1972.

(4) From a strict mathematical point of view, these "density functions" are actually "measures" or "distributions" rather than functions. Thus on p.4, the notation "db" and "dc" should have followed equations (2.4), for total rigor. Historically, this issue has not turned out to be of major importance; see the discussion at the end of section (v), and the more rigorous discussion in Box, George E.P. and Jenkins, Gwilym M, Time-Series Analysis: Forecasting and Control, Holden-Day, San Francisco, Calif., 1970, p.274-283.

(5) More precisely, classical maximum likelihood theory specifies a unique log likelihood measure of goodness of fit, for the simple model above, including the normal noise distribution as part of the model. This measure of fit is a special case of what we will describe in more detail in section (vi), based upon the concepts of section (v). From a conservative Bayesian point of view, this measure is taken to be the logarithm of the probability of truth of a model, conditional upon the observed data, assuming a prior probability distribution which is "flat" when described in terms of the coefficients of the model as written, and relying on the space of these coefficients and of the data as encoded to provide us with the measures over which these probability distributions are defined.

In a sense, this criterion provides a meaningful estimate of the relative probability of truth of the coefficient values considered, subject to the constraint that the model is assumed to be "true," for some value of the coefficients, in whatever sense it is possible for a statistical model to be "true." One of our primary objectives in this thesis is to point out tangible, correctible deficiencies in the classical idea of looking for "truth" as such, in statistical dynamic models; in Chapter (V), we will point out that verbal models and statistical models are subject to similar difficulties, in basic matters. (Statistical models in a very hard science, such as pure physics, may be different, however.) In section (vii), where the new alternatives are discussed on a theoretical level, we have been careful to emphasize that these approaches to the practical prediction of time-series can be understood as an offshoot of the more general and more coherent Bayesian philosophy of induction, as briefly sketched in section (v). The maintenance of this connection is especially important to the social sciences, where the Bayesian framework has many applications beyond those of explicit data analysis; see, for example, Raiffa, Howard, Decision Analysis: Introductory Lectures on Making Choices Under Uncertainty, Addison-Wesley, Reading, Mass., 1968.

(6) Once again, our discussion has been based on the maximum likelihood point of view, which will be called into question in section (vii). For those who are concerned with predictive power, and not with the likelihood of truth as such, the success of various noise models depends less on their "truth" in the process at hand and more on the robustness of the associated estimation procedure; thus, one may choose to regard the regression procedure described above as an independent algorithm, which can be derived from maximum likelihood theory but which is still a distinct object able to stand alone. From this point of view, then, the methods above do not require an assumption of a normal distribution.

(7) Ezekiel, Mordecai and Fox, Karl A., Methods of Correlation and Regression Analysis, Wiley, New York, 1959, Third Edition, chapter six.

(8) Postrikov, Foundations of Galois Theory, Pergamon Press, McMillan, New York, 1962, p. vii.

(9) Wasan, M.T., Parametric Estimation, McGraw-Hill, New York, p.161-162. Alternative techniques exist, but the ones listed by Wasan are second-order - they require the calculation of second derivatives, which are more numerous than first derivatives and may be expensive to calculate. The Marquadt algorithm, the better-known alternative, assumes that the likelihood function is quadratic, an assumption we will not make in this thesis, in later sections; also, it incurs heavy costs in other ways. In the Appendix, we have proposed a procedure for handling the convergence difficulties cited by Wasan, for variations on the theme of steepest descent; in the case of multivariate ARMA(1,1) estimation, at least, resulting convergence times have been reasonable.

(10) Dixon, W.J. BMD Biomedical Computer Programs: X-Series Supplement, U. of California Press, Berkeley, Calif., June 1972, p.177.

(11) Brode, John, Time-Series-Processor - CSP, available from Project Cambridge, MIT, 5th floor, 575 Technology Square, Cambridge, Mass. 02139. A manual for the revised version of TSP may be forthcoming in the MIT Press.

(12) National Bureau for Economic Research, TROLL/1 Primer, available c/o 9th floor, 575 Technology Square, Cambridge, Mass. 02139.

(13) Friedman, Milton, A Theory of the Consumption Function, Princeton University Press, Princeton, N.J., 1957, p.20-31. Friedman's discussion here leaves open somewhat the question of how permanent income is determined; the simplified model we use as our example assumes a simple exponential learning process, based on actual income.

(14) The most fundamental source for this point of view is Carnap, Rudolf and Jeffreys, Richard C., Studies in Inductive Logic and Probability, U. of California Press, Berkeley, Calif. 1971. Discussions of its direct applications to statistics may be found in Box and Jenkins, op. cit. (note 4), p.250-252, and in Kendall, M.G. and Stuart, A., The Advanced Theory of Statistics, Hafner Publishing Co., NY, Second Edition, 1970, Vol.2, p.150.

(15) However, in the philosophy of science, there is occasional reference to the "cosmological principle" that we expect p(data) for the observed data to end up reasonably large, once a full spectrum of theories has been studies. In other words, one expects that the observed data will not be an unusual local coincidence, according to a "true" theory; one expects that data, as observed from earth, in particular, are likely to be typical of data observed elsewhere. Such an additional assumption would not be necessary, or even logical, if we felt that we had p(model) available for the full range of possible models, along with p(data model). However, when we ask about the probability that new models, as yet unformulated, may turn out to be valid, the "cosmological principle" does have something to tell us.

(16) Immanual Kant, A Critique of Pure Reason

(17) Carnap and Jeffreys, op. cit. (note 14).

(18) See the references of note 14. Also see Wasan, op.cit., (note 9) p.150-152; Anderson, R.L. and Bancroft, T.A., Statistical Theory in Research, McGraw-Hill, New York, 1952, p.101; Hays, William L., Statistics for the Social Sciences, Holt Rinehart and Winston, Second Edition, 1973, p.841-842 and 816-821.

(19) Hays, op. cit. (note 18), later chapters; Lindley, D.V., "Professor Hogben's Crisis - a Survey of the Foundations of Statistics", Applied Statistics, Vol.7, No.3, 1958,p.186-198; Raiffa, H. and Schlaifer, R., Applied Statistical Decision Theory, chapter thirteen; Hogg and Craig, Introduction to Mathematical Statistics, McMillan, London, Third Edition,

p.208-209; Wasan, op. cit. (note 9), p.184, definition 5, and subsequent discussions. (Also Box, G.E. and Tiao, G., book coming out.) The two computer programs generally available in Cambridge for Bayesian estimation are described in Brode, John, op. cit. (note 11) and in Schlaifer, R., User's Guide to the AQD Programs, Part III, p. 18, available c/o the Harvard Business School.

(20) In this area, too, the traditional formulation by Carnap and Jeffreys is under question. Shimony, Abner, "Scientific Inference", in Nature and Foundation of Scientific Theories, Colodny, ed., U. of Pittsburgh Press, especially p.100; Solomonoff, "Mathematical Foundations of Induction", manuscript available in 1964 at the MIT Artificial Intelligence Laboratory, from Prof. Minsky; Barker, Stephen F., "The Role of Simplicity in Explanation", in Current Issues in the Philosophy of Science, Feigl and Maxwell, eds., Holt Rinehart and Winston, 1961.

(21) Alpert, Marc and Raiffa, Howard, "A Progress Report on the Training of Probability Assessors," available in 1971 as an unpublished manuscript from the office of Prof. Raiffa in the Littauer Building, Harvard U.

(22) Box and Jenkins, op. cit. (note 4), p. 274. Note that the "Bayesian estimates" suggested on p.252 of this reference, and also the approximations suggested on p.277, involve disregarding this term, with or without small adjustments elsewhere.

(23) Mosteller, C. Frederick and Rourke, Robert E., Sturdy Statistics: Nonparametric and Order Statistics, Addison-Wesley, Reading, Mass., 1973; Tukey, John W., "A Survey of Sampling From Contaminated Distributions", in Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Olkin, Ingram and G. Sudhish, Wassily Hoeffding, William G, Madow, Henry B. Mann, eds., Stanford U. Press, Stanford, Calif., 1960, p. 448-485.

(24) Box and Jenkins, op. cit. (note 4),p.121-124. Our formula here is a special case.

(25) Cochrane, D. and Orcutt, G.H., "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms", Journal of the American Statistical Association, March 1949, p.34.

(26) Forrester, Jay W., Industrial Dynamics, MIT Press, Cambridge, Mass., 1961.

(27) Forrester, Jay W., World Dynamics, Wright-Allen Press, Cambridge, Mass., 1971; the successor to this book was The Limits to Growth, by Meadows, D.H. and D.L, Randers, J., and Behrins, W., Signet Books, New York, 1972, Third Printing; the lack of empirical validation, and other important aspects of this work are discussed in Models of Doom: A Critique of The Limits to Growth, by Cole, H.S.D., Freeman, Christopher, Pavitt, K.L.R., and Jahoda, Marie, Universe Books, New York, 1973. A few alternative approaches are sketched in Chapter (V).

(28) Hogg and Craig, op. cit. (note 19), p. 250-253.

(29) The "ideal types" idea originated with Max Weber; a review of the early idea may be found in Max Weber's Ideal Type Theory, by Rogers, Rolf E., Philosophical Library Inc., New York, 1969.

(30) Feierabend, Ivo K. and Rosalind L., and Gurr, T., eds., Anger, Violence and Politics, Prentice-Hall, Englewood, N.J., 1972, p.114-118.

(31) Jacobson and Mayne, Differential Dynamic Programming, American Elsevier, NY, 1970, especially chapter six.

(32) Samuelson, op. cit. (note 3), p. 345, begins a discussion of this topic; his own attitude is more partial to the traditional Keynesian approach, but his discussion clearly indicates that automatic adjustment factors have been of great interest to some economists.

(33) Jacobson and Mayne, op. cit. (note 31).

(34) Bryson, Arthur E. Jr. and Ho, Yu-Chi, Applied Optimal Control, Ginn and Company, Waltham, Mass., 1969, p.361.

(III) THE MULTIVARIATE ARMA(1,1) MODEL: ITS SIGNIFICANCE AND ITS ESTIMATION

(i) INTRODUCTION

In recent years, the "ARMA" model for statistical processes has become very popular both in industry and in certain parts of the social science community. This popularity is due partly to the landmark book by Box and Jenkins, Time-Series Analysis(1), which places emphasis on the application of ARMA models to predicting future values of time-series variables. Using their approach, one fits a separate model to each variable of interest, a model of the variable as a mixed "AutoRegressive Moving-Average" process of some very complex order; one uses these models, variable by variable, to make predictions of the future.

Our emphasis here is quite different. Our concern, from the beginning, was with studying the interaction between different variables - national assimilation and communications, as described in Chapter (VI) - rather than with the prediction of time-series in isolation from each other; univariate studies were carried out only to help us evaluate methods for dealing with the more general case. With "causal analysis" of this sort, where many variables must be considered together, multiple regression still is the most popular technique by far.(2). Nevertheless, one of the theorems of

Box and Jenkins - that the presence of errors in data-collection can turn a simple regression process into an ARMA process - can be generalized very easily to the multivariate case, as we will show below. Thus one might think of multivariate ARMA analysis as generating the same coefficients as a multiple regression analysis, but "corrected" for the effects of measurement errors. In Chapters (IV) through (VI), we will discuss the empirical work which has convinced us that such measurement noise is not only common, but may also have a drastic effect in reducing the quality of predictions of a model fit by ordinary regression.

In practice, there are two difficulties with using the generalization of the ARMA model to the multivariate case. First, and most important, is the sheer computational difficulty of estimating a full, multivariate ARMA model (a "vector ARMA" model). Hannan(3), in 1970, described the current techniques in this area as follows:"Though there are, no doubt, circumstances in which a vector mixed moving-average autoregressive model will give a much better fit with a given number of constants than either a moving-average or autoregressive model (i.e. ordinary regression - PJW), the computational complications are so great that it can be doubted whether the more complicated model will be used, and we do not feel that the techniques at the present stage are sufficiently near to being practically useful to be included in this book." In 1973,

Prof. George E. Box, of Box and Jenkins, mentioned to us that the heavy orientation of his own work to the univariate case was due in large part to such difficulties. In section (iii) of this chapter, we will describe how we have been able to apply the dynamic feedback algorithm, described in Chapter (II), to overcome this difficulty; in section (iv), we will describe in detail the computer routine, now available to social scientists from Hawaii to London, which we have written to use this method in estimating multivariate ARMA processes.

Second, and more persistent, is the difficulty of "too many degrees of freedom" with ARMA models. If, in addition to accounting for many variables at once, we also added "higher-order" terms, as discussed by Box and Jenkins, the number of parameters in these models could become hopelessly large; such higher-order models could be estimated by a variant of our algorithm below, but the substantive value of the estimates would be questionable. In practice, however, our interest in the ARMA model does not lie in its capacity for being made ever and ever more complicated; our interest is in the possibility of accounting rationally for the problem of "measurement noise," the problem of errors in measuring and indexing the underlying variables which one is trying to study. Thus we will restrict ourselves here to considering "white noise" (random noise, uncorrelated with itself across time but possibly correlated from

variable to variable at the same time) in the process of measurement; all of the ARMA models we will discuss are of the variety which Box and Jenkins would call "ARMA(1,1)" models. While the degrees of freedom problem makes it impractical in most cases to consider more complex, more realistic models of measurement noise, it has been our hope that the difference between accounting for measurement noise and not accounting for it at all would be enough to overcome most of the problems of real-world prediction. However, as we pointed out in Chapter (I), this hope has only been partly realized; given the impracticality of adding too many degrees of freedom to these models, our current opinion is that the "robust method" described in sections (vii) and (xi) of Chapter (II) will be crucial to any further progress with the real-world problems.

Let us now define more precisely what we mean by an "ARMA(1,1)" model. Box and Jenkins(4) define an ARMA(1,1) process "z_t" as a stationary process governed by the scalar equation:

$$z_t = \phi z_{t-1} + a_t - \theta_1 a_{t-1}, \qquad (3.1)$$

where "a_t" is a random normal noise process of covariance σ_a^2, and where "t" is the time period. (We recall from Chapter (II) that "random" means that the process has no correlation with itself across time, or with other processes in the system at earlier times.) Also, note that we are treating "t" as a subscript

here solely because this is the way it appears in Box and Jenkins.

By contrast, the classic autoregressive model may be written, in the univariate case:

$$x_t = \phi x_{t-1} + a_t. \qquad (3.2)$$

The term "a_t" in this equation refers to "process noise," to a random impulse which will affect the true value of x at time t, and thereby affect later values of x as well. In practice, however, the measured data, which we may call "z_t", may differ from the true values of the variable we are trying to study, which we may call "x_t". The difference between the true value and the measured value, $z_t - x_t$, may be called, "measurement noise." If we postulate that this measurement noise, like the random impulses which govern x_t itself, is a random process, then we arrive at the following modification of the classic regression model (3.2):

$$\begin{aligned} x_t &= \phi x_{t-1} + b_t \\ z_t &= x_t + c_t, \end{aligned} \qquad (3.3)$$

where b_t and c_t both are normal random normal noise processes with zero means and with no correlation between each other.

Box and Jenkins(5) have pointed out that any stationary process, z_t, such as the z_t of (3.3) or (3.1), may be completely characterized by knowledge of its correlation function (or, more precisely, its autocovariance), Z_T, across time:

$$Z_T \triangleq E(z_t z_{t-T}).$$

This equation states that "Z_T" is defined to equal $E(z_t z_{t-T})$; "$E(z_t z_{t-T})$", in turn, is the notation we will use to indicate the "expected value" or "mean value" across all times t, of the product, $z_t z_{t-T}$, throughout the statistical process under study. Box and Jenkins argue(6), in the context of discussing general processes which include (3.3) as a special case, that the autocorrelation function of the "z_t" in (3.3) has the same characteristics as that of "z_t" in (3.1); therefore, they conclude that the former statistical process, as a generator of "z_t", is equivalent to a process of the second kind. In other words, a "z_t" generated by a process such as (3.3) will appear to obey a phenomenological equation such as (3.1).

More precisely, Box and Jenkins ask us to consider the following process, in connection with (3.3):

$$\begin{aligned} w_t &= z_t - \phi z_{t-1} = (x_t + c_t) - \phi(x_{t-1} + c_{t-1}) \\ &= (\phi x_{t-1} + b_t + c_t) - \phi x_{t-1} - \phi c_{t-1} \\ &= b_t + c_t - \phi c_{t-1} \end{aligned} \qquad (3.4)$$

From the randomness of b_t and c_t, it is clear that the autocorrelation of this process will be zero for time intervals, T, larger than one. From this information, and from the Gaussian character of the process, they conclude immediately that w_t is itself a simple

moving average process of order one, representable as:

$$w_t = a_t - \theta_1 a_{t-1},$$

for some θ_1 and some random process a_t; if we recall the definition of w_t in (3.4), and substitute, we find that our z_t from (3.3) obeys an equation representable as (3.1). Those readers who have difficulty with this equivalence should refer back to Box and Jenkins.

In the social sciences, however, most dynamic processes of interest involve more than one variable. Fortunately, it is easy to generalize the definitions and results above to the multivariate case, by treating sets of variables as vectors. Thus we can define a multivariate ARMA(1,1) process as a process which obeys:

$$\vec{z}_t = \vec{a}_t + \theta\vec{z}_{t-1} + P\vec{a}_{t-1}, \quad (\theta \text{ and } P \text{ matrices}) \qquad (3.5)$$

where "$\vec{a}_t$" is a vector random process, obeying a multivariate normal distribution(7):

$$p(\vec{a}_t) = \frac{1}{\sqrt{(2\pi)^n \det A}} \exp(-\tfrac{1}{2}\vec{a}_t^T A^{-1} \vec{a}_t), \qquad (3.6)$$

where A is the covariance matrix of this process, where the off-diagonal terms of A allow us to account for the possibility of cross-correlations in the noise process, and where n is the dimensionality of the vectors $\vec{z}$ and $\vec{a}$. We can also define a noisy

time-series regression process as one obeying:

$$\vec{x}_t = \Theta\vec{x}_{t-1} + \vec{b}_t$$
$$\vec{z}_t = \vec{x}_t + \vec{c}_t, \qquad (3.7)$$

where "$\vec{b}_t$" and "$\vec{c}_t$" are random normal processes of dimensionality n, as was "$\vec{a}_t$" above, with covariance matrices which we will call B and C. If we define "$\vec{w}_t$" as $\vec{z}_t - \Theta\vec{z}_{t-1}$, then the rest of (3.4) goes through exactly as before, yielding a process with zero autocorrelation for $T>1$, representable, as before, as a simple moving average process, i.e. as $\vec{a}_t + P\vec{a}_{t-1}$; thus Box and Jenkins' argument for equivalence goes through in its entirety, with equal validity, in the multivariate case.

Our main concern, in this chapter will be with the estimation of the coefficients in (3.5) and (3.6), for a given set of data, $\{\vec{z}_t\}$. However, since our interest in (3.5) and (3.6) comes from our interest in (3.7), it is of interest to see how we could go back, <u>after</u> fitting a model of the form (3.5), to derive the coefficients of an equivalent model of the form (3.7). In the following section, we will elaborate on the mathematical details of this process. For the social scientist, however, the most interesting conclusion from this argument will be the equivalence between "Θ" in (3.5) and "Θ" in (3.7). Thus, the θ_{ij} estimated by the ARMA estimation program itself may be thought of as "corrected" regression coefficients. Just as the usual regression coefficient, b_{ij}, is called a "b coefficient" or "beta coefficient", our "corrected regression

coefficient," Θ_{ij}, may be called a "theta coefficient"; in a similar way, our P_{ij} may be called a "rho coefficient."

(ii) THE RECONSTRUCTION OF A WHITE NOISE MODEL FROM A VECTOR ARMA MODEL

After fitting a vector ARMA(1,1) model, a model of the form (3.5) and (3.6), how do we derive the coefficients of the equivalent model of the form (3.7), assuming that an equivalent model does exist? Recalling that Gaussian stationary processes are completely characterized by their autocovariance functions(8), we may phrase this question as follows. For a given value of Θ, P and A, in (3.5) and (3.6), we wish to find values for Θ, B and C in (3.7), such that the autocovariance matrices Z_n will be the same for the two processes, for all time increments n. Let us use the notation "Θ" to represent the "Θ" in equation (3.5), and "Φ" to represent the "Θ" in (3.7); these two matrices will turn out to be equivalent to each other, but for now we must establish the equality.

The autocovariance matrix, Z_n, is defined as being made up of components, $Z_{n,ij}$, defined as follows:

$$Z_{n,ij} \triangleq E(z_{t,i} z_{t-n,j}),$$

where $z_{t,i}$ refers to the value of the i-th component of the vector $\vec{z}_t$, the value of the vector $\vec{z}$ at time t; from another point of view, $z_{t,i}$ may be regarded as the value of the individual variable z_i

at time t. Note that we have used the same notation here as in section (i) to define the autocovariance. From this definition, we may immediately deduce that:

$$Z_{n,ij} \triangleq E(z_{t,i}z_{t-n,j}) = E(z_{t-n,j}z_{t,i})$$

$$= E(z_{t,j}z_{t+n,i}) \triangleq Z_{-n,ji}$$

$$Z_n = Z_{-n}^T \quad \text{for all } n.$$

Thus if $Z_n^{(a)} = Z_n^{(b)}$, for our two processes "a" and "b", for $n \geq 0$, then the equality will hold for $n \leq 0$, and vice-versa; thus we need only consider $n \geq 0$ in determining equivalence.

From the randomness of "$\vec{a}_t$", "$\vec{b}_t$" and "$\vec{c}_t$", from the causal structure of our equations, and from our assumptions about the normal distributions governing these processes, we have:

$$E(a_{t,i}a_{t+n,j}) = \delta_{n0}A_{ij} = \delta_{n0}A_{ji} \qquad (3.8)$$

$$E(b_{t,i}b_{t+n,j}) = \delta_{n0}B_{ij} = \delta_{n0}B_{ji} \qquad (3.9)$$

$$E(c_{t,i}c_{t+n,j}) = \delta_{n0}C_{ij} = \delta_{n0}C_{ji} \qquad (3.10)$$

$$E(a_{t,i}z_{t+n,j}) = 0 \quad \text{for} \quad n < 0 \qquad (3.11)$$

$$E(b_{t,i}z_{t+n,j}) = 0 \quad \text{for} \quad n < 0 \qquad (3.12)$$

$$E(c_{t,i}z_{t+n,j}) = 0 \quad \text{for} \quad n < 0 \qquad (3.13)$$

$$E(b_{t,i}x_{t+n,j}) = 0 \quad \text{for} \quad n < 0 \qquad (3.14)$$

$$E(c_{t,i}x_{t+n,j}) = 0 \qquad (3.15)$$

$$E(b_{t,i}c_{t+n,j}) = 0 \qquad (3.16)$$

Note that "δ_{ij}", the Kronecker delta, is defined to equal 1 if i=j, and zero otherwise.

Let us begin by calculating the autocovariance matrices, Z_n, for an ARMA(1,1) process as in (3.5). To make our calculations more explicit, let us transform (3.5) into:

$$a_{t,i} = z_{t,i} - \sum_j \theta_{ij}z_{t-1,j} - \sum_j P_{ij}a_{t-1,j}. \qquad (3.17)$$

Let us multiply (3.17) on the right by $a_{t,k}$, and take the expectation of the resulting equation on both sides:

$$E(a_{t,i}a_{t,k}) = E(z_{t,i}a_{t,k}) - \sum_j \theta_{ij}E(z_{t-1,j}a_{t,k}) - \sum_j P_{ij}E(a_{t-1,j}a_{t,k}),$$

which, by (3.8) and (3.11), reduces to:

$$E(z_{t,i}a_{t,k}) = A_{ik} = A_{ki} \qquad (3.18)$$

Multiplying (3.17) by $a_{t-1,k}$, and taking expectations, we get:

$$E(a_{t,i}a_{t-1,k}) = E(z_{t,i}a_{t-1,k}) - \sum_j \theta_{ij}E(z_{t-1,j}a_{t-1,k}) - \sum_j P_{ij}E(a_{t-1,j}a_{t-1,k}),$$

which, by (3.8) and (3.18), reduces to:

$$E(z_{t,i}a_{t-1,k}) = (\theta + P)A \qquad (3.19)$$

Now let us multiply (3.17) on the right by $a_{t-n,k}$, for n greater than one:

$$E(a_{t,i}a_{t-n,k}) = E(z_{t,i}a_{t-n,k}) - \sum_j \theta_{ij}E(z_{t-1,j}a_{t-n,k}) - \sum_j P_{ij}E(a_{t-1,j}a_{t-n,k}),$$

which by (3.8), and by changing the arbitrary origin of our expectation notation, reduces to:

$$E(z_{t,i}a_{t-n,k}) = \theta E(z_{t,j}a_{t-n+1,k}),$$

which, by induction starting from (3.19), gives us:

$$E(z_{t,i}a_{t-n,k}) = \theta^{n-1}(\theta + P)A \qquad n > 0 \qquad (3.20)$$

Now let us multiply (3.17) on the right by $z_{t-n,k}$, for n greater than one:

$$E(a_{t,i}z_{t-n,k}) = E(z_{t,i}z_{t-n,k}) - \theta E(z_{t-1,j}z_{t-n,k}) - PE(a_{t-1,j}z_{t-n,k}),$$

which, by (3.8) and by change of origin in expectation, reduces to:

$$Z_n = \theta Z_{n-1} \qquad n > 1$$
$$Z_n = \theta^{n-1}Z_1 \qquad n \geq 1 \qquad (3.21)$$

Multiplying (3.17) by $z_{t-1,k}$, we get:

$$E(a_{t,i}z_{t-1,k}) = E(z_{t,i}z_{t-1,k}) - \theta\, E(z_{t-1,j}z_{t-1,k}) - P\, E(a_{t-1,j}z_{t-1,k}),$$

which, by (3.11) and (3.18), reduces to:

$$Z_1 = \Theta Z_0 + PA \qquad (3.22)$$

Multiplying (3.17) by $z_{t,j}$, we get:

$$E(a_{t,i}z_{t,j}) = E(z_{t,i}z_{t,j}) - \Theta\, E(z_{t-1,j}z_{t,k}) - P\, E(a_{t-1,j}z_{t,k}),$$

which, by (3.18), (3.19) and our definitions, reduces to:

$$A = Z_0 - \Theta(Z_1^T) - P\,((\Theta + P)A)^T,$$

which by substitution from (3.22) reduces to:

$$A = Z_0 - \Theta Z_0 \Theta^T - \Theta A P^T - PA\Theta^T - PAP^T \qquad (3.23)$$

Equations (3.21) and (3.22) are clearly enough to determine the Z_n for n greater than zero, given A, P, Θ and Z_0.

From the stationarity of the process (3.5), we know that the true variance matrix of $\{\vec{z}_t\}$ does exist, and must satisfy (3.23), just as it is consistent with all the equations from which (3.23) has originally been derived. From the stationarity of (3.5) and (3.7), we may also deduce that both $\emptyset$ and Θ have the property that there exist no nonzero matrices M such that $M = \Theta M \Theta^T$ or $M = \emptyset M \emptyset^T$. (9). It is worth noting, however, that this property, which we will make use of here, involves a much weaker assumption than that of stationarity.(10). At any rate, from this property, we may deduce that the solution, Z_0, to (3.23), is unique; if there had been two

distinct solutions, Z_0 and Z_0', then $M = Z_0 - Z_0'$ would be a nonzero matrix violating our assumption for Θ. At any rate, given our restriction on Θ, equations (3.21), (3.22) and (3.23) are sufficient to define the matrices Z_n as functions of Θ, P and A.

Now let us calculate the autocovariance matrices as functions of B, C and $\emptyset$. Let us rewrite (3.7):

$$b_{t,i} = x_{t,i} - \sum_j \emptyset_{ij} x_{t-1,j} \qquad (3.24)$$

$$c_{t,i} = z_{t,i} - x_{t,i} \qquad (3.25)$$

Let us multiply (3.24) by $b_{t,k}$ and take expectations:

$$E(b_{t,i}b_{t,k}) = E(x_{t,i}b_{t,k}) - \emptyset\, E(x_{t-1,j}b_{t,k}),$$

which, by (3.10) and (3.15), reduces to:

$$E(x_{t,i}b_{t,k}) = B_{ij} = B_{ji} \qquad (3.26)$$

Let us multiply (3.24) by $b_{t-n,k}$, for n greater than zero:

$$E(b_{t,i}b_{t-n,k}) = E(x_{t,i}b_{t-n,k}) - \emptyset\, E(x_{t-1,j}b_{t-n,k}),$$

which, by (3.9), by changes of time origin and by induction, reduces to:

$$E(x_{t,i}b_{t-n,k}) = \emptyset^n\, E(x_{t,i}b_{t,k}) = \emptyset^n B, \qquad (3.27)$$

where the last step comes from substituting (3.26).

Now let us multiply (3.24) by $x_{t-n,k}$, for n greater than zero:

$$E(b_{t,i}x_{t-n,k}) = E(x_{t,i}x_{t-n,k}) - \emptyset\, E(x_{t-1,j}x_{t-n,k}),$$

which, by (3.14) and by induction, reduces to:

$$E(x_{t,i}x_{t-n,k}) = \phi^n E(x_{t,i}x_{t,k}) = \phi^n X_0 \qquad (3.28)$$

Multiplying (3.24) by $x_{t,k}$, we get:

$$E(b_{t,i}x_{t,k}) = E(x_{t,i}x_{t,k}) - \phi\, E(x_{t-1,j}x_{t,k}),$$

which by (3.26) and (3.28) reduces to:

$$X_0 = \phi\,(\phi X_0)^T + B = \phi X_0 \phi^T + B \qquad (3.29)$$

Now let us shift to considering (3.25), multiplying it by $x_{t-n,k}$, for n greater than or equal to zero:

$$E(c_{t,i}x_{t-n,k}) = E(z_{t,i}x_{t-n,k}) - E(x_{t,i}x_{t-n,k}),$$

which, by (3.15) and (3.28), reduces to:

$$E(z_{t,i}x_{t-n,k}) = X_n = \phi^n X_0. \qquad (3.30)$$

Multiplying (3.25) by $c_{t,k}$, we get:

$$E(c_{t,i}c_{t,k}) = E(z_{t,i}c_{t,k}) - E(x_{t,i}c_{t,k}),$$

which, by (3.10) and (3.15), leads to:

$$E(z_{t,i}c_{t,k}) = B_{ik} = B_{ki} \qquad (3.31)$$

Multiplying (3.25) by $z_{t,k}$, we get:

$$E(c_{t,i}z_{t,k}) = E(z_{t,i}z_{t,k}) - E(x_{t,i}z_{t,k}),$$

which by (3.31) and (3.30) gives us:

$$Z_0 = C + X_0 \qquad (3.32)$$

Multiplying (3.25) by $z_{t-n,k}$, for n greater than zero, we get:

$$E(c_{t,i}z_{t-n,k}) = E(z_{t,i}z_{t-n,k}) - E(x_{t,i}z_{t-n,i}),$$

which by (3.13) and (3.30) reduces to:

$$Z_n = \phi^n X_0 \qquad n>0 \qquad (3.33)$$

Now: our problem is to find ϕ, B and C given θ, A and P. Assuming that the Z_n are nonsingular, equations (3.33) and (3.21) clearly tell us:

$$\phi = \theta \qquad (3.34)$$

To find B, let us begin by left-multiplying equation (3.22) by $\theta^{-1}=\phi^{-1}$:

$$\theta^{-1}Z_1 = Z_0 + \theta^{-1}PA.$$

From (3.33) and (3.34), this reduces to:

$$X_0 = Z_0 + \theta^{-1}PA$$

Let us left-multiply this by θ, right-multiply it by θ^T, and subtract the results from the original equation:

$$X_0 - \theta X_0\theta^T = Z_0 - \theta Z_0\theta^T + \theta^{-1}PA - PA\theta^T$$

Substituting in from (3.23) and (3.29), we get:

$$B = A + \theta AP^T + PA\theta^T + PAP^T + \theta^{-1}PA - PA\theta^T$$

$$= A + \theta AP^T + PAP^T + \theta^{-1}PA \qquad (3.35)$$

To find C, let us left-multiply (3.32) by θ, right-multiply

the result by θ^T, and subtract the result from (3.32) proper:

$$Z_0 = \theta Z_0 \theta^T = C - \theta C \theta^T + X_0 - \theta X_0 \theta^T,$$

which by (3.29) reduces to:

$$Z_0 - \theta Z_0 \theta^T = C - \theta C \theta^T + B,$$

which by (3.35) and (3.23) gives us:

$$A + \theta A P^T + P A \theta^T + P A P^T = Z_0 - \theta Z_0 \theta^T$$

$$= C - \theta C \theta^T + (A \pm \theta A P^T + P A P^T + \theta^{-1} P A),$$

which reduces to:

$$P A \theta^T = C - \theta C \theta^T + \theta^{-1} P A,$$

which can be solved by:

$$C = - \theta^{-1} P A. \qquad (3.36)$$

As with equation (3.23), our assumptions about θ lead to uniqueness in this solution, in the same way.

In summary, equations (3.36), (3.35) and (3.34) give us the values of $\emptyset$, B and C necessary to the construction of a model of the form (3.7), equivalent to a process known to fit (3.5) and (3.6) for a given set of coefficients θ, P and A. These values, however, may yet be insufficient; in other words, there may be no values of C, B and $\emptyset$ able to yield an equivalence. Box and Jenkins' argument, cited earlier, states that processes of the form (3.7) will always have equivalents of the form (3.5); they did not state the converse.

If Θ, P and A, in equation (3.36), should require that "C" not be a positive symmetric matrix - what a variance matrix is supposed to be - then we may conclude that our estimates of Θ, P and A contradict the hypothesis that the process at hand fits a model of the form (3.7). For the purpose of actual social science modelling, equation (3.34) tells us that "Θ" coming out of ARMA estimation can not only be used in forecasting, but can also be treated as a description of the underlying social dynamics, $\emptyset$; therefore, we have decided, in our statistical programs, to concentrate on the task of estimating this Θ matrix, and the other ARMA coefficients, rather than adding routines to operationalize (3.35) and (3.36). The terms "beta coefficient" and "b coefficient" are already widely used, in describing the matrix elements of ordinary regression; therefore, our computer routines call the Θ_{ij} "theta coefficients", to emphasize the parallel with regression. The "P_{ij}" are called "rho coefficients," and the "A_{ij}" are simply called "error covariance."

(iii) THE ESTIMATION OF MULTIVARIATE ARMA PROCESSES

Now let us move on to the central question of this chapter: the estimation of multivariate ARMA(1,1) processes. As we pointed out in section (i), it seems rather clear that techniques for multivariate ARMA estimation have not, in the past(11), been reduced to a computational cost approaching that of classical regression. One might have imagined that some kind of spectral techniques might exist in parts of the literature which Hannan and Box and Jenkins are unaware of. However, Jenkins is co-author of one of the classic textbooks on the application of spectral methods to statistics, and has been fully aware of such recent developments as the fast Fourier transform(12); Hannan's book(13) also indicates a full awareness of the possibilities for spectral analysis.

Box and Jenkins, in their Time-Series Analysis, do present(14) a technique for the estimation of ARMA processes. This technique, while described in univariate terms, is phrased in such a way that it extends very easily to the multivariate case; we will find, however, that the extension involves costly computations. They begin with the maximum likelihood technique, as described in Chapter (II); in other words, they set themselves the task of maximizing:

$$L = \log p(\text{observations} \mid \text{model}),$$

They note, on p.210, that the "a_t" in equation (3.1) can be calculated as functions of θ_1, ϕ and a_1, by use of the equation itself, and that the $\{a_t\}$ contain all the information we have available about $\{z_t\}$, given θ_1, ϕ and a_1; thus they put:

$$L = \log p(a_1 \ldots a_T \mid \theta_1, \phi, a_1^*, \sigma_a^2),$$

where "T" is the last time-period for which data are available. Given that a_t is a normally distributed random variable, they get:

$$\begin{aligned} L(\theta_1,\phi,a_1^*,\sigma_a^2) &= \log \prod_{t=1}^{T} p(a_t \mid \theta_1,\phi,a_1^*,\sigma_a^2) \\ &= \log \prod_{t=1}^{T} \left(\frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\tfrac{1}{2}\frac{a_t^2}{\sigma_a^2}\right)\right) \qquad (3.37) \\ &= k - T \log \sigma_a - \frac{1}{2\sigma_a^2}\sum_t a_t^2 \end{aligned}$$

(k is a constant)

In the multivariate case, we need only use the multivariate normal distribution, (3.6), for $\vec{a}_t$, to get an equivalent expression:

$$L(\theta,P,A,\vec{a}_1^*) = k - \frac{T}{2} \log \det A - \tfrac{1}{2}\sum_t \vec{a}_t^T A^{-1} \vec{a}_t \qquad (3.38)$$

They note, in (3.37), that the term "$\sum a_t^2$" is a function only of of θ_1, ϕ and a_1^*, not of σ_a^2, and that we can go on to minimize this term without consideration of σ_a^2. Also, they have a rather elaborate discussion, on p.211 and 502, about finding "good" values for a_1^*, or for estimates of prior data used to predict a_1^*; our own

interpretation is that the "best" empirical value for a_1^* is simply the value of maximum likelihood(15) for the data given, and that the best procedure is simply to append a_1^* to the list of parameters to be estimated. In any case, as Box and Jenkins point out, the choice of procedure here should make little difference, for long or moderate time-series.

At this point, with two parameters to estimate - θ_1 and ϕ - Box and Jenkins suggest that the parameters be **lumped** into a coefficient vector, $\vec{B}$, to be analyzed by the general method of "nonlinear estimation", discussed verbally on p.231 and defined specifically on p.504 as the Marquardt algorithm. The first step of this algorithm is to construct some initial estimate, $\vec{B}_0$, of the coefficient vector, $\vec{B}$. The second step is to calculate the a_t and L for this value, $\vec{B}_0$, by using the dynamic equation, (3.1). With T periods of time t, this implies on the order of T calculations; if we use the multivariate dynamic equation, (3.5), with n variables and Cn terms to be calculated and added per variable, this implies Cn^2 calculations per time period, Cn^2T calculations in all. ("C" will be used throughout this discussion as an arbitrary proportionality constant.) The third step is to calculate the derivatives:

$$\frac{\partial a_t}{\partial B_k} \text{ , for all k and t,}$$

by differentiating (3.1) to get an iterative equation:

$$\frac{\partial a_t}{\partial B_k} = \theta_1 \frac{\partial a_{t-1}}{\partial B_k} + a_{t-1} \frac{\partial \theta_1}{\partial B_k} - z_{t-1} \frac{\partial \phi}{\partial B_k} \qquad (3.39)$$

(From a formal point of view, their differentiation is straightforward, because the "a_t", "a_{t-1}", θ_1 and ϕ are all considered here to be functions of $\vec{B}$; all the differentiations are carried out with respect to this vector, $\vec{B}$. The set of observed data, $\{z_t\}$, is a constant parameter throughout this entire analysis.)

In the multivariate case, we must recall that any coefficient B_k may affect any component, $a_{t,i}$, of the error vector, indirectly, and we will see that the iterative equation below for calculating these derivatives does not allow us to limit our attention to, say, $\frac{\partial}{\partial B_k}(\vec{a}_t^T A^{-1} \vec{a}_t)$; thus the generalization of Box and Jenkins' method requires us to compute:

$$\frac{\partial a_{t,i}}{\partial B_k} \text{ , for all k, i and t,}$$

by using the iterative rule which comes from differentiating (3.5) in the same way as we differentiated (3.1) above:

$$\frac{\partial a_{t,i}}{\partial B_k} = \sum_j \left(\frac{\partial P_{ij}}{\partial B_k}\right) a_{t-1,j} - \sum_j P_{ij} \left(\frac{\partial a_{t-1,j}}{\partial B_k}\right) - \sum_j \left(\frac{\partial \theta_{ij}}{\partial B_k}\right) z_{t-1,j} \qquad (3.40)$$

For each actual B_k in Θ or P or elsewhere, this equation will still require Cn^2 calculations for any period of time t, to handle all the possible combinations of i and j. Thus with Cn^2 coefficients B_k, and Cn^2 calculations per time period per coefficient, and T time periods, this leads to a grand total of Cn^4T calculations. And this is only the beginning.

The next step, in the general nonlinear estimation routine, as discussed by Box and Jenkins on p.232, is to go back to our likelihood function, (3.37) or (3.38), and substitute in a first-order Taylor series for a_t or $\vec{a}_t$ in terms of $\vec{B}$. In the multivariate case, this gives us the major term:

$$\sum_{i,j}\sum_{t}\Big(a_{t,i}-\sum_{k}\Big(\frac{\partial a_{t,i}}{\partial B_k}(B_k-B_k^{(0)})\Big)A_{ij}^{-1}\Big(a_{t,j}-\sum_{m}\frac{\partial a_{t,j}}{\partial B_m}(B_m-B_m^{(0)})\Big)\Big),$$

leading to a generalized form of the matrix which Box and Jenkins unfortunately call "A":

$$\sum_{i,j,t}\frac{\partial a_{t,i}}{\partial B_k}\,A_{ij}^{-1}\,\frac{\partial a_{t,j}}{\partial B_m}$$

For the Cn^4 combinations of k and m, the calculation of this matrix requires the summation of n^2T terms per combination, and Cn^6T calculations in all. By summing the products of the two terms on the

right, over j, for all i, t and m, we may reduce the cost down to Cn^5T. But at this point, the simplifications stop; an "M" of these dimensions, with these properties, is clearly central to the algorithm presented in Box and Jenkins. We could go on to discuss further details of the Marquardt algorithm in the multivariate case, but the number of calculations required - Cn^5T - is already large enough to contrast strongly with the new algorithm we will present below.

Now: how do we arrive at a less expensive algorithm to accomplish the same objectives?

To begin with, we will build our new algorithm on a well established foundation, the classic method of steepest ascent; we will maximize $L(\Theta,P,\vec{a}_1,A)$ by writing:

$$B_k^{(n+1)} = B_k^{(n)} + w\, g_k \frac{\partial L}{\partial B_k}, \qquad (3.41)$$

where w is an arbitrary scale factor to be adjusted during maximization, and where g_k is an arbitrary positive "metric factor" to be applied to B_k. We will include Θ, P and $\vec{a}_1$ as components of $\vec{B}$; however, <u>we will not include A</u>. Starting from a given $\vec{B}^{(0)}$, $A^{(0)}$ and w, we will first compute $\frac{\partial L}{\partial B_k}$. Then we will compute $\vec{B}'$. From $\vec{B}=\vec{B}'$ alone, equation (3.5) allows us to compute all the $\{\vec{a}_t\}$, from times t=1 to t=T. It is a well-known fact, for a <u>given</u> set of data, $\{\vec{a}_t\}$, that the maximum likelihood estimate "A" of the covariance matrix generating this data as a random process of zero mean, will simply be

the observed covariance of the $\{\vec{a}_t\}$:

$$A'_{ij} = \frac{1}{T} \sum_t a'_{t,i} a'_{t,j} .$$

Thus for a given B', we can maximize the likelihood function (3.38) by finding the $\{\vec{a}'_t\}$, and picking A' accordingly. For this combination, we will immediately be able to estimate $L(A', \vec{B}')$, by equation (3.38). If $L(A', \vec{B}')$ is less than $L(A^{(0)}, \vec{B}^{(0)})$, we may reduce w in half and try again. Eventually, for w small enough, we may be sure that $L(A', \vec{B}') > L(A^{(0)}, \vec{B}') > L(A^{(0)}, \vec{B}^{(0)})$, if $\frac{\partial L}{\partial B_k}(A^{(0)}, \vec{B}^{(0)}) \neq 0$, by the definition of the derivative. We may then set $\vec{B}'$ to be the new $\vec{B}^{(0)}$, and A' to be the new $A^{(0)}$. As a practical matter, if $L(A', \vec{B}') \gg L(A^{(0)}, \vec{B}^{(0)})$, we may increase the value of w, to speed convergence. Also, while it would complicate the logic above to change g_k while changing w, it would not hurt to choose a new value g_k while estimating $\frac{\partial L}{\partial B_k}$.

At any rate, this procedure clearly allows a steady improvement in our choice of A and $\vec{B}$, up until a local maximum is attained - i.e. until $\frac{\partial L}{\partial B_k} \approx 0$. The steepest ascent method, like other variational algorithms, including the Marquardt algorithm, does not have the capacity to insure that local maxima are also global maxima. In principle, this means that supplementary routines of varying complexity may be added to the basic algorithm. In practice, we will follow Box and Jenkins by placing emphasis on reasonable initial estimates of $\vec{B}_0$; we will discuss this, and the practical problem

of speeding up convergence, in section (iv).

We face one real theoretical problem in converting the steepest ascent method into a useful algorithm for ARMA estimation - how to calculate the derivatives $\frac{\partial L}{\partial B_k}$ at an acceptable cost. The most elegant way to solve this problem is by the direct application of the ordered derivative concept mentioned in Chapter (II), or at least to apply related concepts; however, in order to avoid the use of unfamiliar mathematics, and in order to make our derivation self-contained, we will use a more conservative, algebraic derivation here.

Let us begin by recalling that "$a_{t,i}$" in equation (3.5) can be considered to be a function of $\vec{a}_1$, θ and P, insofar as (3.5) allows us to solve for the $\vec{a}_t$ by repeated application. In fact, it is simple for us to write out this solution explicitly, for times t greater than one:

$$\vec{a}_t = \sum_{m=2}^{t} P^{(t-m)}(\vec{z}_m + \theta\vec{z}_{m-1}) + P^{t-1}\vec{a}_1 \qquad (3.42)$$

The validity and uniqueness of this solution can be proven easily by induction; for t=2, this expression reduces to (3.5) with t=2; for t+1, equation (3.5) gives us:

$$\vec{a}_{t+1} = \vec{z}_{t+1} + \theta\vec{z}_{(t+1)-1} + P\vec{a}_t,$$

and, if we substitute in from (3.42) for $\vec{a}_t$, as the induction

hypothesis allows us to, we get:

$$\vec{a}_{t+1} = \vec{z}_{t+1} + \theta\vec{z}_{(t+1)-1} + P\left(\sum_{m=2}^{t} P^{t-m}(\vec{z}_m + \theta\vec{z}_{m-1}) + P^{t-1}\vec{a}_1\right)$$

$$= \vec{z}_{t+1} + \theta\vec{z}_{(t+1)-1} + \sum_{m=2}^{t} P^{t+1-m}(\vec{z}_m + \theta\vec{z}_{m-1}) + P^{t}\vec{a}_1,$$

which reduces to (3.42) for the case t+1. Thus, if we think of "$\vec{a}_t$" as a shorthand for the algebraic expression (3.42), it is clear that we can make use of (3.40) above, and of the similar equation which comes from differentiating (3.38):

$$\frac{\partial L}{\partial B_k} = \frac{\partial}{\partial B_k}\left(-\tfrac{1}{2}\sum_{i,j,t} a_{t,i}A^{-1}_{ij}a_{t,j}\right)$$

$$= -\tfrac{1}{2}\sum_{i,j,t}\left(\frac{\partial a_{t,i}}{\partial B_k}\right)A^{-1}_{ij}a_{t,j} - \tfrac{1}{2}\sum_{i,j,t} a_{t,i}A^{-1}_{ij}\left(\frac{\partial a_{t,j}}{\partial B_k}\right),$$

which, by the symmetry of A, equals:

$$\frac{\partial L}{\partial B_k} = -\sum_{i,j,t}\left(\frac{\partial a_{t,i}}{\partial B_k}\right)A^{-1}_{ij}a_{t,j} \qquad (3.43)$$

It will also be convenient for us to write out (3.40) explicitly for the cases $B_k=\theta_{rs}$, $B_k=P_{rs}$ and $B_k=a_{1,r}$:

$$\frac{\partial a_{t,i}}{\partial \theta_{rs}} = -\sum_j P_{ij}\left(\frac{\partial a_{t-1,j}}{\partial \theta_{rs}}\right) - \delta_{ir}z_{t-1,s} \qquad (3.44)$$

$$\frac{\partial a_{t,i}}{\partial P_{rs}} = -\sum_j P_{ij}\left(\frac{\partial a_{t-1,j}}{\partial P_{rs}}\right) - \delta_{ir}a_{t-1,s} \qquad (3.45)$$

$$\frac{\partial a_{t,i}}{\partial a_{1,r}} = -\sum_j P_{ij}\left(\frac{\partial a_{t-1,j}}{\partial a_{1,r}}\right) \qquad (3.46)$$

Now let us define a new variable, $w_{t,i}$, by induction:

$$w_{T,i} = -\sum_j A^{-1}_{ij} a_{T,j}$$

$$w_{t,i} = -\sum_j A^{-1}_{ij} a_{t,j} - \sum_j w_{t+1,j} P_{ji} \qquad (3.47)$$

(Note that $w_{t,i}$ is really $\frac{\partial^+ L}{\partial a_{t,i}}$, in the notation of Chapter (II).)

We now claim that:

$$\frac{\partial L}{\partial \theta_{rs}} = -\sum_{t=2}^{T} w_{t,r} z_{t-1,s} \qquad (3.48)$$

(In Chapter (II), this would follow from a direct application of the chain rule for ordered derivatives, equation (2.13), without reference to (3.44) or (3.45).) This claim can be proven, by proving the more general proposition which follows, and by considering the special case m=1:

$$\frac{\partial}{\partial \theta_{rs}} \left(-\tfrac{1}{2} \sum_{t=m}^{T} \vec{a}_t^T A^{-1} \vec{a}_t \right) = -\sum_{t=m+1}^{T} w_{t,r} z_{t-1,s} + \sum_i w_{m,i} \frac{\partial a_{m,i}}{\partial \theta_{rs}} \qquad 1 \le m \le T$$

(For m=1, recall that $\vec{a}_1$ is an externally supplied parameter, not affected by θ. In the case m=T, the first sum of the right will be held simply to have zero terms.) Differentiating the term on the left, as we did with (3.43), we can simplify this new proposition:

$$-\sum_{t=m}^{T} \sum_{i,j} \frac{\partial a_{t,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{t,j} = -\sum_{t=m+1}^{T} w_{t,r} z_{t-1,s} + \sum_i w_{m,i} \frac{\partial a_{m,i}}{\partial \theta_{rs}} \qquad 1 \le m \le T \qquad (3.49)$$

We will prove this by induction on m, downwards from the case m=T. For m=T, this expression reduces to:

$$-\sum_{i,j} \frac{\partial a_{T,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{T,j} = \sum_{i} w_{T,i} \frac{\partial a_{T,i}}{\partial \theta_{rs}} ;$$

substituting in for $w_{T,i}$ from (3.47), this is equivalent to:

$$-\sum_{i}\sum_{j} \frac{\partial a_{T,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{T,j} = \sum_{i} \left(-\sum_{j} A^{-1}_{ij} a_{T,j} \right) \left(\frac{\partial a_{T,i}}{\partial \theta_{rs}} \right) ,$$

which clearly holds true. Now we try to prove (3.49) for $T > m \geq 1$, on the assumption, provided by the induction hypothesis, that it is true for m+1. We note, for $m < T$, that:

$$-\sum_{t=m}^{T}\sum_{i,j} \frac{\partial a_{t,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{t,j} = -\sum_{t=m+1}^{T}\sum_{i,j} \frac{\partial a_{t,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{t,j} - \sum_{i,j} \frac{\partial a_{m,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{m,j} ,$$

which by the induction hypothesis is equivalent to:

$$-\sum_{t=m}^{T}\sum_{i,j} \frac{\partial a_{t,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{t,j} = -\sum_{t=m+2}^{T} w_{t,r} z_{t-1,s} + \sum_{i} w_{m+1,i} \frac{\partial a_{m+1,i}}{\partial \theta_{rs}} - \sum_{i}\sum_{j} \frac{\partial a_{m,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{m,j},$$

which, by substitution from (3.44), equals:

$$-\sum_{t=m+2}^{T} w_{t,r} z_{t-1,s} + \sum_{i} w_{m+1,i} \left(-\delta_{ir} z_{m,s} - \sum_{j} P_{ij} \frac{\partial a_{m,j}}{\partial \theta_{rs}} \right) - \sum_{i}\sum_{j} \frac{\partial a_{m,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{m,j},$$

$$= -\sum_{t=m+2}^{T} w_{t,r} z_{t-1,s} - w_{m+1,r} z_{m,s} - \sum_i \sum_j w_{m+1,i} P_{ij} \frac{\partial a_{m,j}}{\partial \theta_{rs}}$$

$$- \sum_i \sum_j \frac{\partial a_{m,i}}{\partial \theta_{rs}} A^{-1}_{ij} a_{m,j}$$

$$= -\sum_{t=m+1}^{T} w_{t,r} z_{t-1,s} + \sum_i \left(\frac{\partial a_{m,i}}{\partial \theta_{rs}}\right) \sum_j \left(-P_{ij} w_{m+1,i} - A^{-1}_{ij} a_{m,j}\right),$$

which, by (3.47), equals:

$$-\sum_{t=m+1}^{T} w_{t,r} z_{t-1,s} + \sum_i \frac{\partial a_{m,i}}{\partial \theta_{rs}} w_{m,i} ,$$

proving (3.49), as required in induction, for the case m; with the induction complete, (3.49) is proven, and the special case (3.48) follows immediately.

In a similar way, we claim that:

$$\frac{\partial L}{\partial P_{rs}} = -\sum_{t=2}^{T} w_{t,r} a_{t-1,s} \quad ; \qquad (3.50)$$

the proof of this claim is exactly the same as that of (3.48), above, if we replace all instances of "θ_{rs}" by "P_{rs}", of "z" by "a", and of references to (3.44) by references to (3.45).

Our final claim is that:

$$\frac{\partial L}{\partial a_{1,i}} = w_{1,i} \qquad (3.51)$$

This comes out as a special case of:

$$\frac{\partial}{\partial a_{1,r}} \left(-\tfrac{1}{2} \sum_{t=m}^{T} \vec{a}_t^T A^{-1} \vec{a}_t \right) = -\sum_{t=m}^{T} \sum_{i,j} \frac{\partial a_{t,i}}{\partial a_{1,r}} A^{-1}_{ij} a_{t,j}$$

$$= \sum_{i} w_{m,i} \frac{\partial a_{m,i}}{\partial a_{1,r}} \quad ,$$

for m=1. This follows from induction, too, by the exact same proof used for (3.49), but with "θ_{rs}" replaces by "$a_{1,r}$", with references to (3.44) replaced by references to (3.46), and with references to "z_{t-1}" replaced by zeroes... i.e. with these terms left out.

In short, equations (3.49), (3.50) and (3.51) will give us all the derivatives we need, to operationalize (3.41), once we have computed the "$w_{t,i}$" in equation (3.47). Equations (3.49), (3.50) and (3.47) each require us to carry out Cn^2 computations, for every period of time t, and (3.51) requires us to carry out fewer. Thus the total number of computations required, to get all of the derivatives, is Cn^2T per iteration. This is substantially less than the Cn^5T per iteration of the Box and Jenkins method; for an "n" (number of variables) of about ten, it implies a thousand-fold reduction of cost. Also, we may recall that it requires Cn^2T iterations even to solve for the $\vec{a}_t$, given $\{\vec{z}_t\}$, $A^{(0)}$ and $\vec{B}^{(0)}$; thus the cost of our method here is on the order of the theoretical minimum. Even classical regression costs on the order of Cn^2T operations per analysis(16); thus the technique above brings

ARMA(1,1) analysis down into the range of costs acceptable to those who now can afford multiple regression.

(iv) DESCRIPTION OF COMPUTER ROUTINE TO ESTIMATE ARMA PROCESSES

Our primary goal, in applying the dynamic feedback method to the problem of ARMA estimation, was to construct an operating computer program for use with social science data. This program was written as a new command, "ARMA", in the "TSP" (Time-Series Processor) package for economists, which in turn is a major subsystem of the MIT Cambridge Project Consistent System for social scientists; through TSP, the program has been available for several months to anyone with access to the MIT Multics machine (built by Honeywell), which, as part of the ARPA computer network, can be used directly from all types of computer consoles in a variety of cities from Honolulu, to Washington D.C., to London, England. Donald Sylvan, working with Prof. Bobrow at the University of Minnesota, has made extensive use of this routine to evaluate the impact of American aid programs overseas. The usage of this program is documented in the current TSP manual(17); our concern in this section is with the mathematics behind the program.

In order to convert the algorithm of section (iii) into a working computer program, it was necessary for us to go back and deal

with a number of more practical issues.

To begin with, how do we choose the values for "w" and "g_k" in (3.41) to give us enough progress per iteration to make the reductions in cost we have cited meaningful? Wasan(18) has pointed to this difficulty as the central problem in using steepest ascent in ordinary problems of statistical estimation. We did encounter this difficulty in some of our earlier tests, but quickly found a simple interpretation of the problem and a solution.

In essence, the problem is one of scaling. Suppose that we have two variables - say, world population and average births per female - to be called "z_1" and "z_2", respectively, and to be used in predicting each other's future values. Let us suppose that about 10% of the value of each variable can be explained by the value of the other variable in the preceding year. Then the maximum likelihood value for Θ_{12}, for our data, will be a number in the billions; Θ_{12}, when multiplied by a "z_2" which is much less than one, must lead to a product, $\Theta_{12}z_2$, on the order of billions. Θ_{21}, by similar logic, must be on the order of billionths. A change on the order of unity in Θ_{12} will have very little effect on L, because it represents such a small fraction of the current value of Θ_{12} or of z_1. Thus $\frac{\partial L}{\partial \Theta_{12}}$ will be extremely small, even if Θ_{12} has been misestimated by, say, 10%. On the other hand, a very small change in Θ_{21}, much less than unity, could still double the value of Θ_{21}, and thus lead to a very

large effect on L; thus $\frac{\partial L}{\partial \theta_{21}}$ will be a very large number, if θ_{21} has been misestimated by, say, 10% or so. Looking at (3.41), we can see what the result would be without the "g_k" terms: θ_{12}, which requires a huge change in absolute terms, would be changed very little, while θ_{21}, which requires a small change, would be changed by a much larger amount. Balanced improvement in the two coefficients would be impossible. One might imagine the possibility of imbalanced growth - that "w" might be made very small at first, that θ_{21} would converge to its own optimum, where it would generate a zero derivative, and that "w" could then grow enough to allow θ_{12} to move to <u>its</u> optimum. However, in general, the coefficients in a statistical model are not so completely independent of each other. If the optimum of θ_{21} depends at all on the estimate of θ_{12}, then our first small changes in θ_{12} will lead to enormous derivatives from θ_{21} again, destabilizing the system again before there is a chance for "w" to build up enough to allow a large increase in θ_{12}. Thus, at least when the scaling problem is severe, the hope of imbalanced growth is not an answer to the danger of slow convergence.

The solution of this problem was rather straightforward for ARMA estimation; we simply scaled the variables of the problem according to a common scale. More precisely, we achieved the same effect by setting:

$$g_k \text{ for } \theta_{rs} = \frac{\sigma_r^2}{\sigma_s^2}\ ;\quad g_k \text{ for } P_{rs} = \frac{\sigma_r^2}{A_{nn}}\ ,$$

where "σ" refers to the standard deviation. On a more sophisticated level, what we are doing here is trying to maximize the expected progress per iteration, in light of our prior probabilistic knowledge about L. We do not expect the units of measurement of the variables to tell us anything about their relative influence on each other; therefore, we demand a choice of "g_k" which insures that a change of units will have no effect on our algorithm. More generally, these variances give us an idea of the expected order of size of a coefficient, and we set "g_k" to keep the changes in line with the expected ratios between sizes and derivatives. To handle the case of $B_k = a_{1,r}$, therefore, we write a rough but reasonable expression:

$$g_r = \max(T(1-P_{rr})A_{rr}, A_{rr})$$

By changing $a_{1,r}$ by a certain amount, we are changing $a_{t,r}$ in units proportional to 1; thus our formula for g_r is like our formula for the g_k with θ_{rs}, except that "σ_s^2" is replaced by "1". If P_{rr} equals one, then this effect will take place on all the $a_{t,r}$, and the analogy is exact. Otherwise, if P_{rr} is smaller, the derivative with respect to $a_{1,r}$ of L will be much smaller, even when the optimal size of $a_{1,r}$ is still just as large; thus we propose a large g_r, in that case. Note, as "$A^{(0)}$" is recalculated in every major iteration, the formulas above encourage us to recalculate the "g_k" at the same time. In Chapters (IV) and (VI) we have included a brief discussion of the success of this general procedure with

the estimations we have carried out; in the Appendix to Chapter (II) we have suggested ways of generalizing the procedure for use with general, nonlinear models.

The choice of "w" requires a similar exercise in prior estimation. At each step, our program looks at three essential pieces of data - $L(A^{(0)},\vec{B}^{(0)})$, $L(A',\vec{B}')$ and $\sum_k (B'_k-B_k^{(0)})\frac{\partial L}{\partial B_k}$. Assuming that L is essentially quadratic, and that the current choice of w is "right" (i.e. that $L(A',\vec{B}')$ is the maximum of the quadratic distribution), the program "expects" that $L(A',\vec{B}')$ will be better than $L(A^{(0)},\vec{B}^{(0)})$ by exactly half what the gradient would appear to indicate. If this expectation is correct, then the program concludes that B' is not only acceptable, but also that there probably is little point in exploring further in the same direction; it sets $\vec{B}^{(0)}=\vec{B}'$, $A^{(0)}$ to A', and begins a new major iteration. If $L(A',\vec{B}')$ is worse, then, by the quadratic assumption, we have overshot by at least a factor of two; w should be cut in half, and a new $\vec{B}'$ tried accordingly. In order to be a bit more conservative have specified in our program that w will be reduced by 40%, if the actual gain is less than 25% of what is indicated by the gradient. If $L(A',\vec{B}')$ is better than 75% of what is indicated by the gradient, our quadratic assumption tells us to double w. In an earlier version, we were more conservative here, and required 100%; however, convergence was slow in some cases, and we reduced the requirement to 87%, which has proven adequate. In the intermediate range, when $\vec{B}'$ is deemed acceptable for

the start of a new major iteration, w is still changed somewhat, for the sake of the next iteration; w is multiplied by twice the actual gain, $L(A',\vec{B}') - L(A^{(0)},\vec{B}^{(0)})$, divided by the gain predicted by the gradient. At the other extreme, if w appears to be far off, the program will multiply w by 4 or by .3 in each minor iteration; more precisely, if w appears to be too small to let us set $\vec{B}^{(0)}=\vec{B}'$, even after w has just been doubled within the same major iteration, or if w appears to be too large after having just been cut to 60%, then a larger change will be tried in the next minor iteration. Flags are set in the program, to force it to stop changing w, as soon as it starts changing w in opposite directions within the same major iteration; in such cases, our procedures above insure that either the last $\vec{B}'$ or the one before it gave an $L(A',\vec{B}')$ much better than $L(A^{(0)},\vec{B}^{(0)})$, and our program will set $\vec{B}^{(0)}$ to this new $\vec{B}'$ for the next major iteration. For reasons similar to those mentioned in the previous paragraph, w is initialized at 1/T.

At each step, the program prints out L, as defined in equation (3.38), and the direction of change of w. After five major iterations, or after L appears to have stabilized to within .01, whichever comes first, the program stops, and asks the user if he wishes to continue; if not, it prints out the analysis so far, and transfers to another program to carry out simulation studies of his model. In the current version described above, five major iterations have usually been enough for a close approximation, for analyses of actual

social science data; for safety, however, we have generally used ten in our own analyses. (Once L is about 0.1 away from its maximum, then the current set of coefficient estimates has almost as high a probability of exact truth - 90% as high - as the maximum likelihood set itself; thus 0.1 is a conservative upper limit to how much accuracy it makes sense to ask for.) Unfortunately, the changes above were made piecemeal over a number of runs on different data, with the final improvements existing only in the basic subroutine incorporated into the MIT version. This routine has a more effective procedure for generating initial estimates than we used with our earliest test data; thus the direct comparison, before and after, would overstate somewhat the relative merits of the current system. In the Appendix to this Chapter, we have provided a numerical example of convergence results before and after these procedures for convergence were introduced.

The fundamental purpose in using ARMA estimation, as we have described it, is to improve upon classical multiple regression. Thus we have decided to use multiple regression itself, to provide the initial estimates, $\vec{B}^{(0)}$. Not only are these likely to be reasonable estimates, in terms of their general order of size and in terms of the size of the biggest terms; they also provide us with an assurance that our ARMA model will either represent an improvement upon multiple regression, or, in some cases, will confirm multiple regression.

Our subroutine has been written to allow other initial estimates, but the main program now available at TSP does not make use of this option. Originally, we used the regression coefficients that come from a standard model including regression constants; our results on Norway, in section (v) of Chapter (VI), were based on that system. With the constants, one introduces a greater degree of freedom into the regression models, to offer a more interesting (though perhaps artificial) comparison against the ARMA models, which do not include that degree of freedom. However, in order to insure convergence under all circumstances, we have eliminated the regression constants in the MIT version; users of that system still have the freedom, in any case, to obtain regression constants based on constant terms by using other modules in the same system or even by another run of the ARMA command. Both the MIT version, and our private version used on the Norway data, print out all the ARMA estimates and regression coefficients, along with the standard deviations of the variables (to assist in interpretation) and the likelihood values for both models. The significance of the various coefficients can be estimated by looking at the likelihood of the models which result when the coefficients are removed from the model.

Several other options have been added, to extend this algorithm somewhat. First of all, there is now provision for "exogenous

variables." In the discussion above, the expression "Θz_{t-1}" could have been replaced systematically by "Θy_{t-1}", where Θ is now a rectangular matrix, and where y_{t-1} includes both z_{t-1} and a few other components; none of the equations above would have had to be changed in form. Our program, in its current form, allows both endogenous and exogenous variables.

Second, there is provision to allow the user to dictate apriori that certain components of Θ will be constrained to equal zero. This is done simply enough, by setting their initial values to zero, and constraining (3.41) to apply only to the other components of Θ. Thus L is maximized as a function of the other components, subject to this constraint. The basic subroutine allows this for any coefficient, but, in the MIT version, we have limited this to those Θ_{ij} for which y_j is an exogenous variable.

Third, there is provision to give the user some ability, at least, to handle nonstationary processes. Box and Jenkins(19) discuss at great length the prominence of nonstationary processes in practical statistics; they point out the value of introducing some kind of careful procedure for dealing with nonstationary processes, even if the procedure must have a less rigorous foundation than the usual statistical processes, in order to give the social scientist confronted with such processes an alternative other than either giving up or using an inappropriate tool. However, the procedures they introduce(20) involve processes which tend to grow

as t^k, for some constant k. By contrast, the commonest processes of growth in the social sciences would appear to be those of exponential growth, processes which may come out of a dynamic relation like that of equation (3.5), but with a choice of "θ" large enough to allow growth. The concept of maximum likelihood, as discussed in Chapter (II), does not require a "θ" that generates a stationary process; thus at first glance, the special procedures suggested by Box and Jenkins might appear irrelevant. However, our estimation procedure has depended on equation (3.6), not just on (3.5) and the likelihood concept. Equation (3.6) implies that the average size of the random component of our process remains the same across time. If we were analyzing a two-hundred-year series of data on the US GNP, for example, this would imply that a $10 billion error in our predictions for 1790 from 1789 should be treated as a smaller matter than an $11 billion error in our predictions of 1973 from 1972; a $10 billion error would always be regarded as less significant than an $11 billion error, <u>regardless</u> of the year in which the error occurred. In practice, the measurement errors and random fluctuations both are likelier to be a fixed percentage of the variable itself - GNP - than to be a fixed independent process. To handle this kind of situation, we have introduced an option, "ARMAWT", to deal with a model of error slightly different from (3.6):

$$p(a_t) = \frac{1}{\sqrt{(2\pi)^H \det A}} \exp\left(-\tfrac{1}{2} \sum_{i,j} \frac{a_{t,i}}{z_{t,i}} A^{-1}_{ij} \frac{a_{t,j}}{z_{t,j}} \right),$$

which is simply the normal distribution for the n-dimensional vector $\begin{Bmatrix} a_{t,i} \\ z_{t,i} \end{Bmatrix}$, and which requires us to calculate A_{ij} as the covariance of this vector. In practice, however, the simulation studies of Chapter (IV) suggest that the ordinary ARMA command generally performs at least as well as ARMAWT, even for most of the nonstationary processes studied.

Finally, provision has been made for the possibility - mentioned in Chapter (II) - that the available data would consist, not of one string of observations across time for our variables, but of a whole set of such strings; a general model of the process of population growth, for example, might encourage us to develop a model for application to data-series involving the same variables across many different countries. In order to handle this possibility, we can use (3.41) as before, but must note:(i) $\frac{\partial L}{\partial \theta_{rs}}$ and $\frac{\partial L}{\partial P_{rs}}$, across all the data strings, will simply equal their sum across all the individual data-strings; (ii) each data-string, S, will require its own $\vec{a}_1^{(S)}$ for initialization; (iii) $\frac{\partial L}{\partial a_{1,r}^{(S)}}$ will, of course, equal the value of $\frac{\partial L}{\partial a_{1,r}}$ calculated in string S. (Note that this example might be a good candidate for the use of "ARMAWT", to prevent the analysis from being dominated by nations of large population... unless such weighting is actually desired.)

APPENDIX: NUMERICAL EXAMPLE OF THE BEHAVIOR OF DIFFERENT CONVERGENCE PROCEDURES

All of the ARMA estimations reported in Chapters (IV) and (VI) were based upon the final form of our algorithm, making use of the special convergence procedures described in section (iv). However, before we installed these procedures, we carried out a number of tests on the preliminary version of the ARMA routine, on simple made-up data sets, in order to check out the accuracy of the routine. One of these simple test series was used before we introduced the possibility of different "g_k" for different "B_k", as in equation (3.41); thus we can see the effect of adding our new convergence procedures by comparing the old test results against a new analysis of the same series. The data series in question is a simple univariate series of length seven - 1.0, 1.2, 1.2, 1.3, 1.5, 1.4, 1.0 - fit to the model $z_{t+1} = \theta z_t + a_t + Pa_{t-1}$. This series does not fit well to a simple arithmetic progression; thus the "distance" from the regression model to the ARMA model turns out to be fairly great; the series is a relatively severe test of convergence possibilities. (The cost per iteration is low, because the series is short, but the progress per iteration in log probability, as a percentage of the gap in log probability, is extremely slow.) Our initial test output

was a string of numbers, which we may arrange in a table:

Major Iteration Number	Theta in $\vec{B}(0)$	LogP $\vec{B}(0)$ at $A(0)$	LogP at $\vec{B}'$, A'	Change of w
1	.8204	-.8925	-.2017	0
"	"	"	1.6138	-1
"	"	"	.2602	-1
2	1.0049	1.6138	1.5624	0
"	"	"	1.6571	-1
"	"	"	.0325	-1
3	.9746	1.6571	1.6773	0
4	.9843	1.6773	1.6849	0
5	.9849	1.6849	1.7004	0
"	"	"	1.7162	1
"	"	"	1.8176	2
"	"	"	2.3374	2
"	"	"	-8.8156	2
"	"	"	3.6485	-1

Table III-1: Example of Convergence Results With Early Version of the ARMA Estimation Routine

"Change of w" means the value of "ntest", a number indexing the source of the current $\vec{B}'$. If the w used in generating $\vec{B}'$ was taken directly from the last major iteration, "0" is used; if w was cut by 40% in the previous minor iteration within the same major iteration, a "-1" is used; and so on, as one might expect from our description in section (iv). Note that the calculations implied by this table include five computations of the gradient of likelihood, and fourteen computations of likelihood (average errors) for sample coefficient vectors $\vec{B}'$.

There follows the script of a TSP session based on the same data-series. The ordinary user, to get into TSP, would have to

sign in on the MIT Multics, then enter the consistent system, and then issue the command "tsp:x" or "tspr:x"; in our own directory, we only needed to issue the command "tspr" directly:

```
tspr                                        (us)

T 23:25    4.916    $ .33                   (tsp)

data$                                       (us)

T 23:25    .965   $.07                      (tsp)

smpl 1 7$ load oldtst$ 1 1.2 1.2 1.3 1.5 1.4 1.0$ end$   (us)

smpl vector                                 (tsp)
  1    7                                    (tsp)

T 23:26    .581   $0.05                     (tsp)

arma oldtst$ end$                           (us)

it.no. 1, from logp=          1.616          ( tsp here on down)
 0;newlogp=       1.709
-1;newlogp=       2.281
it.no. 2, from logp=          2.281
 0;newlogp=     -12.660
-1;newlogp=       1.900
-2;newlogp=       2.864
 1;newlogp=       3.468
 1;newlogp=      -2.119
 1;newlogp=       3.468
it.no. 3, from logp=          3.468
 0;newlogp=     -47.638
-1;newlogp=     -30.132
-2;newlogp=       3.396
-2;newlogp=       4.032
 1;newlogp=       4.706
 1;newlogp=      -0.044
 1;newlogp=       4.706
it.no. 4, from logp=          4.706
 0;newlogp=     -22.527
```

Table III-2: Sample Session With Convergence Results For Final Version of the ARMA Estimation Routine (top section).

```
-1;newlogp=      -11.374                          (tsp)
-2;newlogp=        3.865
-2;newlogp=        4.895
it.no. 5, from logp=          4.895
 0;newlogp=        5.114
 1;newlogp=        5.249
continue?                                         (tsp)
yes                                               (us)
it.no. 1, from logp=          5.249               (tsp)
 0;newlogp=      -11.615
-1;newlogp=       -3.843
-2;newlogp=        4.928
-2;newlogp=        5.381
it.no. 2, from logp=          5.381
 0;newlogp=        5.445
it.no. 3, from logp=          5.445
 0;newlogp=        5.542
it.no. 4, from logp=          5.542
 0;newlogp=        5.204
-1;newlogp=        5.537
-2;newlogp=        5.586
it.no. 5, from logp=          5.586
 0;newlogp=        5.636
 1;newlogp=        5.684
 2;newlogp=        5.846
continue?                                         (tsp)
yes                                               (us)
it.no. 1, from logp=          5.846               (tsp)
 0;newlogp=      -20.319
-1;newlogp=      -12.058
-2;newlogp=        1.427
-2;newlogp=        5.825
-2;newlogp=        5.968
it.no. 2, from logp=          5.968
 0;newlogp=        6.005
it.no. 3, from logp=          6.005
 0;newlogp=        6.021
 1;newlogp=        6.037
 2;newlogp=        6.127
 2;newlogp=        6.020
 2;newlogp=        6.127
```

Table III-2: Sample Session With Convergence Results For Final Version of the ARMA Estimation Routine (middle section)

```
it,no. 4, from logp=      6.127                    (tsp)
 0;newlogp=        .947
-1;newlogp=       4.054
-2;newlogp=       6.046
-2;newlogp=       6.147
it,no. 5, from logp=      6.147
 0;newlogp=       6.162
continue?                                          (tsp)
yes                                                (us)
it,no. 1, from logp=      6.162                    (tsp)
 0;newlogp=       6.184
it,no. 2, from logp=      6.184
 0;newlogp=       6.197
-1;newlogp=       6.203
it,no. 3, from logp=      6.203
 0;newlogp=       6.202
-1;newlogp=       6.219
it,no. 4, from logp=      6.219
 0;newlogp=       6.235
it,no. 5, from logp=      6.235
 0;newlogp=       6.249
continue?                                          (tsp)
no                                                 (us)

logp=     6.249(       1.616)                      (tsp)

predicting oldtst

indep.var.  theta         rho         reg.coeff.    error      rms
oldtst     .9722708     1.636800      .9867076     0.98203e-02 1.276715

reg error = 0.36896e-01                            (tsp)
continue?                                          (tsp)
no                                                 (us)
continue with simulation?                          (tsp)
yes                                                (us)
```

Table III-2: Sample Session With Convergence Results For Final Version of the ARMA Estimation Routine (bottom section)

After this, we went on to check out the simulation routine, and other routines in TSP, to make sure that all was working correctly. Note that the convergence information would not have been printed out

in so much detail if we had turned on the "output some" flag.

Notice that the variance of the error with the ARMA model - .0098203 - was quite a bit smaller than the variance of the error with regression - .036896. This is a good index of the distance between the two models. It is interesting that the economic cost of using ARMA analysis - measured in iterations - would appear to be less when it turns out to be unnecessary, when the distance is small; the number of iterations can get large, mostly in the case where the benefit from using ARMA analysis is also large. We have deliberately used many iterations in this recent run, to confirm that convergence was reasonable after ten major iterations or so, in this difficult special case. Our earlier test run was continued for only five major iterations, as shown in Table III-1; however, in those five iterations, it covered roughly the same distance, in increasing logp, that our new system did in the first two. More significantly, the old routine showed major signs of floundering, and one has the impression that its final breakthrough was partly a matter of luck. The new routine moved systematically towards convergence. Note, also, that the problem of scaling is not unusually great in this case; the variables P and a_1 need scaling vis-a-vis Θ, but with long multivariate time-series one expects a far greater scaling problem, and a more dramatic need for the new procedures suggested in section (iv), and in the Appendix to Chapter (II).

FOOTNOTES TO CHAPTER (III)

(1) Box, George E.P. and Jenkins, G.M., Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, Calif., 1970.

(2) In principle, this statement might call for a statistical survey itself, although for those familiar with the usual procedures in social science it is an understatement. The MIT computing center, which now serves both MIT and Harvard, has put out a brief survey of statistical procedures available to its users on the IBM 370, in its publication AP-77. This survey lists six available statistical packages - the Statistical Package for the Social Sciences, Data-Text, Econometric Software Package, the (IBM) Scientific Subroutine Package, P-STAT (Princeton statistical package), and the BioMeDical package. Five of the six include "multiple regression"; ESP, the sixth, would appear to contain the same provision under the term "simple linear regression." Bayesian statistics are not included in any of the listings. Nonlinear least squares is present only in BMD. (We suspect, however, that ESP - a cousin of TSP - might have this capability by now.) Moving average models are not mentioned. Spectral analysis is not mentioned as such; the X Supplement to the BMD manual indicates its presence in the more recent version of that package, but in none of the others. This probably gives as accurate indication of the dominance of regression analysis in actual work in the social sciences.

(3) Hannan, E.J., Multiple Time-Series, Wiley and Sons, New York, p.394

(4) Box and Jenkins, op. cit., p.76

(5) Box and Jenkins, op.cit., p.30 (bottom of page)

(6) Box and Jenkins, op. cit., p.121-124. Consider the case $q_1=q_2=1$, $d=0$.

(7) Box and Jenkins, op. cit., equation (A7.1.9), p.260.

(8) See note 5.

(9) One way of expressing the idea of stationarity is that the norm of θ^n, and thus of $(\theta^T)^n$, will become arbitrarily small for n sufficiently large; in particular, let us choose an n for which these norms are less than one. If $M=\theta M\theta^{-1}$, then, by substitution and induction, $M=\theta^n M\theta^{Tn}$. If M is nonzero, then

there must exist some vector, $\vec{v}$, of unit length, for which $M\vec{v}$ is of nonzero length; let us pick the unit vector $\vec{v}$ for which the length of $M\vec{v}$ is a maximum. (Given that the matrix M is of finite dimension, at least one such vector must exist.) Then our assumptions clearly tell us that the length of $M\vec{v}$ would be greater than the length of $\Theta^n M \Theta^{Tn}\vec{v}$, contradicting our matrix equality.

(10) Consider, for example, $\Theta = kI$, where k is not +1 or -1, and where I is the identity matrix. If k is larger than +1, this Θ would correspond to a highly nonstationary process. Yet this Θ still meets our requirements. If $M=\Theta M \Theta^T$, then, by substitution, $M=k^2M$; this cannot happen for the "k" we have mentioned, for a nonzero M. Indeed, it would appear that our assumption could only be violated for that infinitesimally small proportion of matrices Θ which have eigenvalues exactly equal to 1 in absolute value. If this were proven, it is conceivable that our reasoning could be extended even to that set of matrices by some sort of limit theorem; however, such possibilities go beyond the scope of our discussion here.

(11) It has been pointed out to us that R.L. Kashyap, in the area of engineering, has suggested a procedure for the "estimation" of models of the form of equation (3.7), in "A New Method of Recursive Estimation in Discrete Linear Systems", in IEEE Transactions, AC-15, #1, p.18-25. "Estimation" in this article is different from what a statistician would call estimation; the article concerns itself with the use of a model of the form (3.7), with coefficients already determined, to predict future values of "z" and the like. Nevertheless, another article by Kashyap in the same journal, "Maximum Likelihood Identification of Stochastic Linear Systems", does present a general method of approach which could be extended to yield an algorithm similar to our own for estimating processes of the form (3.5) above. Kashyap's general algorithm is considerably weaker than our own or Jacobson's, discussed in Chapter (II), insofar as it applies only to linear processes. On page 26, he discusses the possibility of using a representation for his statistical processes involving a "moving average error"; however, he uses a form of moving average with considerably more degrees of freedom than the form used in statistical theory, and comes to the conclusion that such a representation is impossible. A study of the relation between his parameters and ours might yield a solution procedure, like ours discussed in section (ii); alternatively, the same general approach might have been used from the beginning on our own

representation. Wasan (see note 18 below) has pointed out the importance of adding a rational procedure for handling "w" and "g_k", an issue which Kashyap does not discuss, before one can claim to have a workable algorithm in the field of statistics. Kashyap also mentions a notion of "constrained derivative," which looks like a precursor of the "ordered derivative" of Chapter (II), but based upon notions of variational calculus; the concept, as he uses it, does not include his "lambdas" as a set of constrained derivatives, while they correspond very clearly to ordered derivatives in our own system.

(12) Jenkins, G.M. and Watts, D.G., Spectral Analysis, Holden-Day, San Francisco, Calif., 1968. p.313 includes reference to the Cooley-Tukey fast Fourier algorithm, developed in 1965.

(13) Hannan, op. cit., p.32-106, p.127-136, and the greater part of p.245-405.

(14) Box and Jenkins, op. cit.

(15) Strictly speaking, our estimate of $p(a_1)$ should include reference to the general probability of the z_0 which we would deduce, etc. However, as we point out later in the text, our estimation procedures are not very sensitive to the assumption of stationarity; to account for this extra piece of information, z_0, becomes rather doubtful when nonstationarity is involved, and when z_0 represents the beginning of a process previously governed by different dynamics. As in section (vi) of Chapter (II), we have decided that "perfection" on this point would not be worth the cost, especially in light of Box and Jenkins' similar loose approach to the point.

(16) With n dependent variables, and n independent, one must compute two n by n covariance matrices, with each term requiring T multiplications and summations, in conventional regression. When the estimation of such a model is carried out by the separate estimation of n simple regression equations, directly from raw data, one computes an n+1 by n+1 matrix n different times, implying an even greater cost.

(17) The best, most recent description is available in Brode, John, Werbos, Paul and Dunn, Elizabeth, TSP in the Datatran Language, available in draft form from the Cambridge Project, 5th Floor, Technology Square, Cambridge, Mass.; discussions are underway regarding the publication of this manual through the MIT Press. The command language has been changed, to increase flexibility.

(18) Wasan, M.T., Parametric Estimation, McGraw-Hill, New York, 1970, p.151-152.

(19) Box and Jenkins, op. cit., especially p.85-94.

(20) Box and Jenkins, op. cit., p.87 (bottom) and p.113 (top).

(IV) SIMULATION STUDIES OF TECHNIQUES OF TIME-SERIES ANALYSIS

(i) INTRODUCTION

Most of the discussion in this thesis about the disadvantages of multiple regression - the classical mainstay of time-series analysis - has emerged from the study of concrete data in political science. One might ask, however, whether our discussion applies to other sorts of time-series, in economics or ecology or elsewhere. Our verbal discussions, in section (vii) of Chapter (II) and in Chapter (V), suggest that the superiority of the "ARMA" approach and of the "robust" approach are due to special characteristics of the data we have studied; in particular, this superiority may be due to the presence of complex measurement noise. While measurement noise may be almost universal in the social sciences, it would still be very interesting to get some kind of tangible idea about <u>how much</u> measurement noise, of <u>what kind</u>, and <u>where</u>, leads to <u>how big</u> of a failure of ordinary regression. Indeed, in section (vii) of Chapter (II), in discussing the trade-off between the maximum likelihood approach, as represented by ARMA estimation, and the "robust approach", we

emphasized that some weight should be given to the findings of each approach, and that there is no universal prescription for what these weights should be in all cases; even if a universal prescription is impossible, however, a number of clear concrete numerical examples may help us greatly in building up an intuitive map of the tradeoffs.

Simulation studies can provide us with these examples. Indeed, with simulation studies it is possible to generate hundreds of sample time-series, all standardized, all based on known types of statistical process; time-series in the real world rarely offer such tidiness, and rarely allow us to feel so secure in our interpretations. Even the possibility of unique, erratic events can be accounted for, if we insert terms for erratic types of random disturbance into the simulation process, as we will describe below. In principle, one could even simulate unique, all-encompassing shifts, in which one is asked to predict the behavior of a time-series which will, in the future, obey a different system of dynamic laws from those it has obeyed in the past; however, it is not reasonable to expect <u>any</u> statistical routine to

pass this last test, in its most general form. Difficulties of this last sort, in the real world, can only be minimized by intelligent human use of statistics, as we will discuss in Chapter (V).

Our goal in this chapter, then, is to begin the process of mapping out the domains in which different techniques of time-series analysis are appropriate, as indicated by the analysis of simulated data. The territory to be mapped out, in principle, is very vast; it includes all the statistical processes and models, multivariate and nonlinear and highly complex, which could ever be relevant to the social or natural sciences. Thus we have no choice, here, but to try and pick out a subregion of this territory, small enough to be manageable but large enough to illustrate the qualitative factors most important in our discussion.

Our goal, more precisely is to compare the ability of different estimation techniques to fit the coefficients of a simple model, in such a way that it predicts effectively the behavior of an "unknown" process which may actually be more complex than what the model itself can express completely. In social science, in general, we presume that a true, complete

description of the actual processes going on would contain far too many parameters to be estimated from the available data. We will focus on the problem of estimating the simplest model we can think of, of relevance to social science:

$$Z(t+1) = cZ(t) \qquad (4.1)$$

Sample time-series, "Z", of length 200, have been generated by simulating the results of more complex processes; then, for each sample time-series, Z, we have compared the ability of each of our basic estimation techniques to come up with a good value for "c". It should be noted that ordinary linear multivariate estimation problems are simply the extension of this example to the case where Z is a vector and c is a matrix.

In section (iii), we will see that the studies which we have carried out generally support the conclusions outlined in Chapter (I); however, before we can describe these results, it is necessary to define in detail precisely what studies were carried out.

(ii) DEFINITION OF STUDIES CARRIED OUT

The main results of these tests are summarized in Table IV-1, at the end of this chapter; secondary results are summarized in Table IV-2, and the raw computer output is tabulated at length in Tables IV-3 and IV-4. In order to explain precisely what these tables mean, we must define:(i) what the twelve more complex processes are, that we use in simulating sample time-series; (ii) what the six estimation techniques are, that we use to estimate "c" from the first 100 observations of each sample time series;(iii) what the criteria are that we use to evaluate these estimates.

The complex processes used in simulation were all chosen to be "compatible" with (4.1), in the sense that (4.1) could do an adequate job of prediction if the constant "c" were chosen appropriately. This implies a constant average rate of growth for the variable "Z". Thus we decided to focus our attention on twelve processes that generate a single observed time-series, Z, on the basis of homogeneous linear equations.

In our verbal discussion, we have placed great emphasis on the possibility of "measurement noise" or "transient noise", as distinct from "process noise" or

"objective randomness". This has led us to focus on processes which generate an "observed" (or "superficial") variable, Z, as the result of two subprocesses:(i) an "inner", or "objective", process, which determines the evolution of the "true" or "underlying" variable, "X", over time;(ii) an "outer", or "measurement", process, in which Z, the "measured value" of X (or an "index of X"), is determined, by superimposing some noise factor over the true value of X. Z, and only Z, was later made available to the estimation routines. (Strictly speaking,this situation is merely a special case of the more general situation, where one can observe directly only a subset of the variables of dynamic significance.) The first six of these processes we determined by equations of the form:

$$X(t+1) = (1.03)X(t)(1+P(t))$$

$$Z(t) = X(t)(1 + M(t)) \qquad (4.2)$$

"P(t)" and "M(t)" were both "noise processes" of various sorts; "P(t)" represents "objective" or "process noise", while "M(t)" represents "transient" or "measurement noise". Note that we chose a 3% natural

growth rate, per time period, for X and Z; this would seem rather typical for economic and social science data.

The equations generating P(t) and M(t) were different for each of these six processes. In essence, they were chosen from three different noise processes, A, B and C. "A" was a normal random process of mean zero and of variance one:

$$p(A(t)) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(A(t)^2)} \qquad (4.3)$$

Thus "A" is a simple classical noise process, based on a bell-shaped curve. To generate "B", we would first generate a random number, "A", as above. Then, with probability .95, we would set B=A; however, in 5% of the cases, chosen at random, we would set B = 10A. This implies a probability distribution:

$$p(B(t)) = \frac{.95}{\sqrt{2\pi}} e^{-\frac{1}{2}(B(t)^2)} + \frac{.05}{10\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{B(t)}{10}\right)^2} \qquad (4.4)$$

"B" is generated by a "distribution with outliers." To generate "C(t)", a more complex noise process, we would first generate B(t); then we would generate $\theta(t)$, another random variable, by picking $\theta(t)=1$ with

probability 20%, or by setting $\theta(t)$ to a number chosen at random from a uniform distribution between 0.0 and 0.2 in all other cases; we would then generate C(t) via the equation:

$$C(t) = (1- \theta(t))C(t-1) + 4\theta(t)R(t) \qquad (4.5)$$

This procedure for generating C(t) is an attempt to express the idea of noise which "may or may not correlate with itself across time," whose correlation itself, $(1-\theta(t))$, can change randomly with time. Equation (4.5) was further modified by the use of an occasional cutoff, which we will describe below.

The choices used in processes one through six, for insertion into equations (4.2), may be summarized:

Process 1 (and 7): P(t) = .05A ; M(t) = 0

Process 2 (and 8): P(t) = .05A ; M(t) = .15A

Process 3 (and 9): P(t) = .05B; M(t) = .15A

(4.6)

Process 4 (and 10): P(t) = .05A ; M(t) = .15B

Process 5 (and 11): P(t) = .05A ; M(t) = .05C

Process 6 (and 12): P(t) = .05B; M(t) = .05C

Five and fifteen percent errors were chosen on grounds

that they seem "typical"; computer time was not available to replicate this study for different values of these parameters. It should be noted that the appearance of "A" twice with Process 2, above, does not mean that the same random number, A(t), was used in both processes; in general, every time that we needed a random number for a new application, we invoked a new call to "random", the random number generator of the Project Cambridge TSP-CSP system. In equation (4.2), one should also note that an unrealistic change of sign could occur if P(t) should ever equal -1 or less; this would be a very rare event, with the systems we have specified, but even one such incident would persist throughout an entire simulated time-series, making it totally unrealistic as a representative of social-science time-series. Similarly, while measurement errors are occasionally quite gross for social science variables, it is unrealistic to imagine someone getting the sign wrong for such variables as GNP or population. Thus for P(t) and M(t) both we instituted a cutoff of -.75; values less than this were set equal to the cutoff. (No simulations were run without a cutoff; thus it is possible that the cutoff

was never actually invoked.) A more elegant procedure, mathematically, might have been to use e^{P} and e^{M} instead of (1+P) and (1+M) in equation (4.2). However, major and enduring crashes do sometimes occur in those social science variables subject to erratic behavior; for example, phenomena such as zero or negative population seem to be avoided as a result of extraordinary processes different from those operating on normal populations. Thus, on grounds of realism, we decided to use cutoffs instead of a more elegant approach. Also, a cutoff of -15. was used for the value of B(t) inserted into equation (4.5), on grounds that this would prevent the possibility of invoking a cutoff several times in a row on M(t).

The choices above, in brief, give us a chance to look at the four possibilities of no measurement noise, of simple measurement noise, of medium-complex measurement noise, and of complex measurement noise, all in the presence of simple process noise; they also let us look at simple measurement noise and complex measurement noise, in the presence of medium-complex process noise. Processes 1 and 2 closely resemble the processes for which simple regression and ARMA models,

respectively, should be ideal, in theory. The extreme case of zero process noise, the case which should be most favorable to the "robust approach", was not included, on grounds that we are interested in evaluating that approach under more normal, mixed conditions. In order to account for the possibility of more complex processes, without violating homogeneity, we introduced Processes 7 through 12 based on the following equations, which yield a growth rate of 1.6% (from the linearized difference equation) :

$$X(t+1)=(.38X(t)+.35X(t-1)+.3X(t-2))(1+P(t)) \qquad (4.7)$$

$$Z(t) = X(t)(1 + M(t))$$

The choices of P(t) and M(t) here were identical to those with Processes 1 through 6, as indicated in equations (4.6), and the same cutoffs were used.

For each of the twelve processes defined above, ten sample time-series, "Z1" through "Z10", were generated. Then, in order to estimate "c" in equation (4.1), for each of those sample time-series, we used the three general techniques discussed throughout this thesis:(i) classical regression; (ii) the "ARMA"

approach; (iii) the "robust approach". The most conventional way of estimating "c", for the model (4.1), is to use a standard regression program to do a maximum likelihood estimation of the related model:

$$Z(t+1) = cZ(t) + k + a(t), \qquad (4.8)$$

where "k" is a constant to be estimated, and "a(t)" is a normal random noise process. In practice, this amounts to doing a least-squares estimation, as in section (ii) of Chaper (II); one hopes that "k", which is expected to be zero, will be estimated as something close to zero. This technique, regression with a constant term, is abbreviated as "reg+k" in our tables.

A better way of using classical regression, to estimate "c" in (4.1), is to use a simpler model, without the meaningless constant term:

$$Z(t+1) = cZ(t) + a(t) \qquad (4.9)$$

This kind of regression can be performed automatically in the Time-Series-Processor system. This technique, regression without a constant term, is abbreviated as "reg" in our tables.

Corresponding to these two simple regression

models are two simple ARMA models. According to our result in the beginning of Chapter (III), the correspondence is more than just one of similarity; the ARMA models below are the generalizations of (4.8) and (4.9) to account for the possibility of "white noise" in the process of data measurement. Model (4.8) corresponds to:

$$Z(t+1) = cZ(t)+a(t)+Pa(t-1)+k, \qquad (4.10)$$

which can be estimated directly in the Project Cambridge Time-Series Processor, by use of the command "ARMA", described in the later part of Chapter (III). This technique for estimating "c" we abbreviate as "arma+k" in the tables at the end of this chapter. Model (4.9) corresponds to:

$$Z(t+1) = cZ(t)+a(t)+Pa(t-1), \qquad (4.11)$$

which can also be estimated by the command "ARMA"; this technique for estimating "c" we will abbreviate as "arma" in the tables at the end of this chapter. Finally, the estimation algorithm described in Chapter (III) allowed us to write another command, ARMAWT, to estimate the model:

$$Z(t+1) = cZ(t)(1+a(t))+Pa(t-1)z(t-1) \quad (4.12)$$

This command, mentioned briefly in Chapter (III), is essentially equivalent to estimating (4.11), with the assumption that the noise process, "a(t)" in (4.11), is determined as a <u>percentage</u> of the actual variable, as in (4.2), rather than a process of constant mean and variance. This technique for estimating "c", we abbreviate as "armawt" in the tables at the end of this chapter.

Finally, we had to find a "robust procedure" for estimating c, drawn from the discussion of section (vii) of Chapter (II). In the pure case of zero process noise, these procedures require us to estimate the initial "underlying" values of the variables measured, and the coefficients of the model, by directly minimizing the average errors of long-term predictions made with this model. In other words, for the estimated initial values, and a given set of coefficients, one makes a full set of predictions for the variables of interest, <u>without ever</u> making use of the measured values of the variables at intermediate times; one uses the average error in <u>these</u> predictions, predictions which are generally long-term predictions,

as one's criterion of fit; one uses the method of steepest descent, or a related procedure, in order to pick coefficients and initial estimates to minimize the total error in <u>these</u> predictions. (A "relaxed" version of these procedures would allow a little bit of allowance to be made for intermediate measured values, in predicting more distant time periods.)

The full multivariate, nonlinear version of this procedure, based on the dynamic feedback algorithm of Chapter (II), was not available at the time these simulations were carried out. However, for equation (4.1), there is a measurement-noise-only model which is much easier to estimate than is usually the case:

$$X(t+1) = cX(t)$$

$$Z(t)=X(t)e^{a(t)} \approx X(t)(1 + a(t)) \qquad (4.13)$$

If the measurement noise, a(t), is on the order of 10% or less, the approximate equality here will be very good, according to the Taylor expansion of $e^{a(t)}$. In order to estimate X(0) and c in this model, one can transform (4.13) algebraically to deduce:

$$Z(t) = X(0)c^{t}e^{a(t)}$$

(4.14)

$$\log Z(t) = t \log c + \log X(0) + a(t)$$

In order to pick the constants, log c and log X(0), to maximize the likelihood of this model, according to standard maximum likelihood theory, one need only perform a simple regression of log Z(t) against the independent variable "t" and a constant term. A special routine to perform this operation, called "GRR" (GRowth Rate), was added to the Project Cambridge Time-Series Processor in January 1974. Note that this routine estimates "c" and X(0) in exactly the same way as the old routine "EXTRAP" did, in the work reported in Chapter (VI).

Finally, after the simulation processes and the estimation techniques are defined, we face the problem of measuring how well the estimation techniques actually perform. We have used two different criteria. First, there is the criterion of predictive power, measured explicitly. For each combination of time-series (out of 12 X 10 = 120 sample time-series) and of estimation technique, we used the value of "c", as estimated for the <u>first</u> 100 time-periods, to try to

predict the values of the variable Z over the remaining 100 periods. For each such set of predictions we calculated four measures of error: (i) r.m.s. (root mean square) average percentage error in predicting periods 101 through 110; (ii) r.m.s. average percentage error in predicting periods 101 through 125; (iii) r.m.s. average percentage error in predicting periods 101 through 150; (iv) r.m.s. average percentage error in predicting periods 101 through 200. (Also, a set of predictions was made, from period 1, to periods 1 through 100.) The exact results of these tests are shown in Table IV-4, for every sample time-series.

Let us define in a bit more detail how these predictions were arrived at. For the regression models, (4.8) and (4.9), we inserted the known value of z(100), and the estimates of c and k, and the most probable value for a(100) (i.e. zero), in order to predict z(101); this prediction for Z(101) was reinserted, along with a(101)=0, to give us a prediction of Z(102), which in turn was reinserted, etc. For equation (4.9), this has the same effect as inserting the estimate of "c" into (4.1), and using (4.1) to make the forecasts. With the ARMA models, equations (4.10), (4.11) and

(4.12), almost the same procedure was used. The measured value of Z(100) was inserted to give the first prediction, the prediction of Z(101); a(100) through a(199) were set to zero. However, the ARMA equations also refer to a(t-1); thus, in the very first round, in predicting Z(101), the value of a(99) has to be accounted for; for a(99), we use the estimated value which was generated by the ARMA estimation procedure which had been used on periods 1 through 100. With equation (4.11), as with equation (4.9), this has the same effect as inserting the estimate of "c" into equation (4.1), starting from the predicted value of Z(101). "Percentage error" was defined, in general, as a percentage of the average of predicted and actual values, on grounds that this is a good intelligible approximation to exponential error in the normal range, and that it does not place overemphasis on outliers. All of these decisions were made, not at the time of simulation, but at the time when the TSP command "ARMA" was written; at the time of simulation, we specified the initial time and the number of periods to predict, and the ARMA (ARMAWT) command carried through the decisions described in this paragraph by itself. (When

the full nonlinear algorithm of Chapter (II) is operationalized, however, we will be able to let the user choose his own index of prediction error, according to what he considers important to policy-makers in his own particular domain of interest. With ARMA, a linear system, it was necessary to make a general choice for all users, based on mathematical rather than substantive considerations.) With the univariate robust approach, there are two reasonable bases for prediction. One is to use (4.14) directly, assuming that a(101) through a(200) equal zero; we abbreviate this method as "ext1". (This corresponds to our old "EXTRAP" procedure, described in Chapter (VI); also, it corresponds to using (4.1), starting from the estimated value of X(0) as the initial value of Z(0).) The other is to use the estimated value of c in (4.1), inserting the measured value of Z(100) into this equation; we abbreviate this as "ext2". The r.m.s. average errors are computed as with the "ARMA" model, automatically, by the command GRR.

Table IV-4 is a bit too complex to be assimilated directly by the intuition. Thus we have summarized the major results of Table IV-4, regarding prediction

errors, in Table IV-2. For each of the twelve simulated processes described in equations (4.2) through (4.7), each of the seven prediction techniques described above, and each of the four prediction intervals, we have calculated the average prediction errors across the ten sample time-series. More precisely, to avoid a picture distorted by outliers, we have tabulated the worst (biggest) of the errors out of the ten, and the average across the remaining nine. The rows containing the average values, for different prediction techniques, are labelled "av"; the rows containing the maximum errors are labelled "max". Also, in column eight, we list the "dispersion" of the errors of the best technique, defined as the average over the nine better sample time-series of the absolute value of the difference between the error in each sample series and the average error.

A quick scan of Table IV-2 indicates a general tendency of "ext2" to be superior substantially to regression; in some cases,"ext2" and "arma" are approximately equal, while in other cases "arma" and regression are approximately equal. A more detailed scan reveals three difficulties with these measures of

predictive power. At long time intervals, errors get so high that it is hard not to worry about the effects of our percentage-taking procedure, and hard to feel fully comfortable about the significance of the averages; this difficulty may not be as real as it seems, but it is worth noticing. A more serious difficulty is the tendency of all prediction techniques to do equally well at very short time intervals, with most of our processes. With short-term predictions, the effects of different estimates of "c" have not had time to build up; thus all of the predictions are close to each other, relative to the very large short-term fluctuations our simulated processes impose. It is the medium and long-term predictions which separate the sheep from the lambs. This reminds us of certain schools of thought in the stock market, who compare the short-term fluctuations of stocks to a roulette game, and who claim that superior analysis makes money only by pointing out longer-term trends. With complex, large-scale multivariate processes in the social sciences, however, one might expect the fluctuations to look a bit smoother through time, even though the measurement noise problem remains. A few of our twelve

processes do show significant differences between estimation techniques in the ten-year prediction tests; these processes may be more representative of the social sciences. A third difficulty is the limitation of having only ten sample time-series per analysis.

In Table IV-1, we have used a second criterion to measure the success of different estimation techniques. We have looked directly at the values of "c", as estimated by the different models. With simulation studies, unlike studies in the real world, we can be sure that the "true" value of c is the same for all the samples of a given process; this is what makes a direct comparison possible. With a direct comparison, one does not worry about having one's conclusions randomized by the effects of random fluctuations in later periods of time, in a limited number of sample time-series; the actual prediction errors in Table IV-2 may be interpreted as a noisy measurement of the quality of the estimates of "c". Indeed, in most studies of political and economic phenomena, people tend to be interested in the validity of the coefficients, "c", and only vaguely aware of the connection between the validity - even in the short-term - and the long-term

predictive power of the resulting model. (This attitude would be quite reasonable when it is a choice between focusing on the validity of "c", or focusing on short-term predictive power. An accurate model of the effects of government policy might reduce prediction error by only 20%, in comparison with a null model, if short-term fluctuations are large enough, in accord with the pattern described in the paragraph above; however, this 20% would include 100% of the effects which the decision-maker can have on the situation.)

For all these reasons, the estimates of "c", evaluated directly for accuracy, appear to be the best criterion to use in evaluating the estimation techniques here. The exact estimates of "c" for each sample time-series are shown in Table IV-3. In Table IV-1, we have summarized this information, for easier interpretation. For each estimation technique, and each simulated process, we have calculated the <u>average</u> value of the estimates of "c", across all ten sample time-series. We have also calculated the "<u>dispersion</u>" of these estimates, the average value of the absolute value of the difference between the estimate of "c" for a given sample time-series and the average estimate

across all ten samples. The rows labelled "av" give the average; the rows labelled "disp" give the dispersion. These calculations were made with the help of a hand calculator, directly from Table IV-4. (Note, however, that the version of Table IV-4 in this chapter has been rounded off, to save space; the calculations were made from the unrounded original.)

Unfortunately, the noise components of our twelve processes, while "unbiased" in the sense of an arithmetic average, do produce a negative shift in the average rate of growth. In order to give some sort of measure of the "true" rate of growth, we have taken the geometric average of the estimates of "c" by "GRR"; this appears in the "av" rows, in column seven, of Table IV-1. Following the logic of section (vii) of Chapter (II), we would contend that the "true average rate of growth" might even be defined as the expected "estimate" or "observation" of the rate of growth, c, based on fitting an exponential curve such as (4.13) implies, over an infinitely long sample of the process in question. (For column seven, we use a data sample ten times as large as that used with any of the specific estimates.) The potential difficulty with the

"robust" technique is not with consistency, the ability to converge to the value most useful in long-term prediction when unlimited data are available, but efficiency, the ability to make full use of the limited data available, as recommended by the maximum likelihood technique. (More precisely,the maximum likelihood method, as sketched out in section (v) of Chapter (II), claims to point to the estimates of maximum probability, conditional upon all information in the observed data.) If simple regression does outperform the robust method, one would expect it to do best for simulated processes which fit a regression model; one would expect the (geometric) average of the estimates of "c" to be equally good for both methods, but one would expect the dispersion to be less with regression, because regression, in exploiting more information per sample of data, can converge more quickly to its asymptotic estimates.

(iii) DESCRIPTION OF RESULTS

In short, in examining Table IV-1, we can sort out two different sources of error in using our estimation techniques:(i) systematic bias, the gap between the

average estimate and the "true" estimate, as indicated in column seven and in the estimates of all the better techniques; (ii) inefficiency, the inability to converge quickly to the asymptotic estimates, as indicated by the dispersion of the estimates across different sample time-series. (In all of what follows, we emphasize that the "true" estimate is being defined as the estimate which leads to the best predictions.) Classic maximum likelihood theory would claim total efficiency as its prime advantage over the robust approach, as discussed above; thus the dispersion errors are of particular interest.

Looking carefully at Table IV-1, we immediately observe a startling fact: in nine out of the twelve simulated processes, the "robust method" outperforms every other method, even in terms of dispersion. Regression without a constant term does better than the robust approach for only two processes, in terms of dispersion : Processes 1 and 7, the simple processes with no measurement noise at all, following a regression model almost exactly; even in these very special cases, the dispersion with the robust method is only slightly larger. Even with Process 1, the ARMAWT

technique outperforms regression by a larger margin than that of regression over the robust approach, which comes in as third. With Process 7, the simple ARMA technique is best. Process 8 is the only other process for which the robust approach is not superior; in that case, where the measurement noise is "white", the situation discussed in section (i) of Chapter (III), the simple ARMA model does a bit better than the robust approach, but both of these two do substantially better than the others. Even with Process 2, where the process and measurement noise are again both "white", the robust approach is ahead. In seven out of the eight remaining processes (all but Process 3), the robust method outperforms all the other methods, except for the simple ARMA models, by at least a factor of two, in all cases.

In summary: even in the domain of statistical efficiency, where the maximum likelihood methods should have their greatest advantage, the robust method enjoys substantial superiority - i.e. dispersion errors less than half the size - in all but the simplest cases, where the advantages of the other methods, where they exist, are slight.

In the domain of systematic bias, where we expect the robust approach to enjoy its greatest advantage, the criteria available are unfortunately less objective. The estimated growth factors, "c", are all less than (1.03) and (1.016), the growth factors inserted into the original sets of processes, due to the expected watering-down effect of random noise. With every one of the twelve processes, however, our estimate of the "true" value of c, in column seven of Table IV-3, is either closer to the original growth factor than are any of the six average estimates, or else within .0002 of whichever of those estimates is closest; this tends to support the value of our estimate in column seven.

Looking at Table IV-1, we see very clearly a strong negative bias, in all the averaged estimates of c, which are from simple regression. In five out of the twelve processes, regression has estimated a <u>negative</u> rate of growth, for processes which we know at least to have a <u>positive</u> rate of growth; thus the very <u>sign</u> of the trends in these processes are reversed. In four of the remaining processes, regression gives a growth rate of less than 1%. <u>The size of these bias errors is much</u>

greater than the average dispersion errors, which, we have noted, were already quite a bit larger than those of the robust method; thus if both sources of error are added together, the overall errors in coefficient estimates are considerably worse than a mere factor of two for regression, in comparison with the robust approach. If we look more closely at the three processes most favorable to regression, in terms of bias error, we find that in two of them the bias error is still larger than the average dispersion error, and that in the third the bias error is still larger than 1%, i.e. larger than 35% of the actual growth rate. The estimates of "c", with a constant term present, are, as one might expect, still worse than those of simple regression. The ARMAWT analysis also performs disappointingly poorly, with negative growth rates for all but four of the processes; in this case, it is theoretically possible that a hidden bug in programming was involved, insofar as cross-checks against existing programs were not possible, but a simple lack of robustness would seem to be a more likely explanation.

The contest between the ARMA and the robust methods is closer, and more interesting. After doing the

analyses of political data, reported in Chapter (VI), we were frankly surprised at how much better the ARMA method did here. In one process - Process 8 - the ARMA model had the same average estimate of "c" as the robust approach did, and a smaller dispersion error; thus, for this one process, the robust approach was actually somewhat inferior to the ARMA approach. On the other hand, as we have noted, process 8 was defined in terms of pure white noise; most social science variables, like the ones studied in Chapter (VI), may be more like processes 11 and 12, or much further in the same direction, in terms of complexity. In four out of the twelve processes, the ARMA and robust approaches gave average estimates of "c" within .0005 of each other; this tends to reinforce the validity of these estimates as an indication of the "true" growth rate. Only for two of the processes was the bias error of the ARMA estimate larger than 1%, relative to the estimate in column seven. In general, the bias errors of the ARMA estimates were less than their dispersion errors. On balance, the robust approach did better, only because the dispersion errors of the ARMA estimates were substantially larger than the errors of the robust

approach for the majority of our processes, especially the more complex processes. (For the twelve simulated processes, in order, the ARMA dispersion errors, as a fraction of the robust method errors in Table IV-1, equalled 1.03, 1.16, 1.54, 2.29, 2.55, 2.13, .84 ,.80, 1.15, 1.21, 4.18 and 1.83.)

In Table IV-1, we have included one other piece of information, of relevance to our discussion in Chapter (III). We have included a description of the number of major iterations required before convergence, with our algorithm for ARMA estimation. In the ARMA estimations, we allowed for ten major iterations before stopping the routine. In the seventh column, in the "disp" rows, we list, first the number of iterations actually required, on the average, before the likelihood scores converged to within 0.1 of their final value. (i.e. The posterior probability of the estimates was at least 90% of the posterior probability of the "most likely" estimates finally converged to.) This average was taken only for those sample time-series in which such convergence was attained before the last iteration. Second, after a colon, we list the number of sample time-series, out of the ten,

in which the 0.1 level of convergence was achieved only on the last iteration or later. In most cases, convergence was achieved well before the last iteration. In those processes where convergence was slower, such as processes 8, 2 and 10, the final estimates of "c" do not appear to have suffered as a result; indeed, the negative bias of the initial estimates obtained from regression was overcome more completely in these processes than in the others. Again, the cost of the ARMA estimation, in terms of iterations, was highest precisely in those cases where the payoff of the approach was also greatest.

Finally, we should say a little about Table IV-2. Here again, the competition is mostly between the robust approach - ext2, more exactly - and the simple ARMA approach. The errors in short-term prediction tend to be watered down and randomized, due to the sheer size of the unpredictable short-term fluctuations, as discussed a few pages back. A closer look at Table IV-3 shows that these prediction errors are affected heavily by outlying time-series. Otherwise, the ARMA technique appears to do a little better here, relatively, than it does with its estimates of "c"; also, the differences

between all the estimation techniques are watered down, with only a few examples of ratios of two in average error. On balance, however, Table IV-2 appears to follow the conclusions for Table IV-1 fairly closely.

	ext	reg+k	arma+k	reg	arma	armawt	(true)
Process av	1.0266	1.0149	1.0151	1.0205	1.0208	1.0211	1.0266
1 disp	.00392	.00692	.00690	.00390	.00404	.00370	1.8:0
Process av	1.0262	.9597	1.0160	1.0035	1.0267	1.0064	1.0262
2 disp	.00480	.01944	.01720	.01090	.00556	.00580	6.6:0
Process av	1.0199	.9181	.9851	.9893	1.0134	.9374	1.0198
3 disp	.00952	.04228	.03626	.01322	.01464	.06372	4.7:0
Process av	1.0270	.9682	1.0203	1.0019	1.0272	.9913	1.0270
4 disp	.00340	.04832	.01556	.02376	.00780	.02638	4.5:2
Process av	1.0274	.9819	.9919	1.0099	1.0156	.9698	1.0274
5 disp	.00540	.04056	.03256	.01876	.01376	.04408	1.6:0
Process av	1.0280	.9935	.9969	1.0174	1.0190	.9714	1.0279
6 disp	.01040	.04620	.04332	.02352	.02220	.04212	6.3:2
Process av	1.0140	.9960	1.0028	1.0110	1.0124	1.0122	1.0140
7 disp	.00200	.00600	.00504	.00180	.00168	.00184	3.1:0
Process av	1.0138	.8590	.9503	.9904	1.0139	1.0042	1.0138
8 disp	.00244	.04760	.05098	.00628	.00194	.00364	7.5:2
Process av	1.0122	.7884	.9198	.9801	1.0106	.9811	1.0122
9 disp	.00396	.08312	.08428	.00792	.00456	.02608	6.0:4
Process av	1.0143	.8967	.9638	.9924	1.0145	.9843	1.0143
10 disp	.00190	.06202	.03932	.01484	.00230	.02218	6.0:5
Process av	1.0153	.9182	.9361	.9947	1.0008	.9399	1.0153
11 disp	.00356	.06494	.05808	.01822	.01488	.05294	2.2:0
Process av	1.0156	.9274	.9547	1.0023	1.0116	.9658	1.0156
12 disp	.00524	.06904	.05542	.01424	.00960	.02808	2.3:0

Table IV-1: Average coefficient estimates, and dispersion errors of estimates, for the six estimation routines and twelve simulated processes defined in section (ii).

Page IV-34

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 1					
10 av	11.7	7.8	8.4	8.2	8.0	7.9	7.4	2.4
""max	29.1	12.2	23.1	23.1	21.6	21.6	21.1	
25 av	13.7	10.2	15.2	15.0	13.4	13.5	12.2	2.5
""max	33.2	18.7	37.7	38.7	33.4	33.5	31.6	
50 av	23.7	18.1	28.4	28.3	24.9	24.8	22.7	6.1
""max	38.8	35.3	66.2	66.2	56.9	57.1	53.2	
100 av	32.1	26.3	57.0	56.8	45.7	45.0	41.7	8.6
""max	53.9	59.4	104.	104.	87.0	88.6	82.2	
			Process 2					
10 av	20	21	32	22	27	18	22	3.1
""max	48	37	58	36	53	31	33	
25 av	22	23	53	33	45	24	32	4.4
""max	55	44	88	61	80	48	55	
50 av	29	29	80	48	69	33	50	7.6
""max	76	60	116	102	106	80	97	
100 av	36	36	120	72	103	40	95	14.5
""max	104	81	149	138	144	105	140	

Table IV-2: Prediction Errors as Defined in Section (ii)

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 3					
10 av	52	25	32	22	34	22	39	5.1
""max	93	51	77	52	62	46	106	
25 av	77	51	61	50	64	46	90	10.6
""max	109	96	112	81	103	84	167	
50 av	98	77	76	78	91	73	122	15.4
""max	139	134	143	114	134	121	184	
100 av	111	97	100	100	116	107	134	22.2
""max	165	161	165	148	165	146	189	
			Process 4					
10 av	30	17	20	17	18	17	24	6.6
""max	42	113	110	33	145	32	100	
25 av	33	22	33	24	27	22	41	3.9
""max	56	119	129	35	172	32	150	
50 av	37	28	57	35	43	26	65	8.7
""max	63	125	149	63	186	49	176	
100 av	44	44	95	62	77	39	101	15.3
""max	91	132	170	119	193	100	188	

Table IV-2: Prediction Errors as Defined in Section (ii)

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 5					
10 av	26	19	19	20	20	20	34	5.1
""max	60	57	78	71	89	69	102	
25 av	29	24	29	28	27	28	58	5.8
""max	61	65	83	82	114	78	141	
50 av	32	29	48	44	38	39	86	9.8
""max	75	82	124	125	156	128	172	
100 av	51	46	90	82	66	61	126	10.8
""max	85	87	154	155	179	161	186	
			Process 6					
10 av	31	22	24	24	27	28	29	8.3
""max	109	43	45	47	56	55	73	
25 av	38	32	41	41	44	44	63	8.2
""max	104	55	85	90	112	111	133	
50 av	57	47	66	67	66	65	95	10.3
""max	119	81	116	123	148	147	166	
100 av	82	70	98	97	93	91	136	13.7
""max	140	116	157	162	174	174	184	

Table IV-2: Prediction Errors as Defined in Section (ii)

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 7					
10 av	7.2	6.7	7.5	5.5	6.9	5.2	5.1	.9
""max	14.7	9.9	15.9	12.5	13.2	10.0	8.4	
25 av	8.0	7.8	12.6	8.8	9.6	6.7	6.6	1.7
""max	17.2	12.4	26.3	21.3	19.0	14.6	13.9	
50 av	12.9	10.6	23.0	15.7	14.1	11.4	11.2	2.8
"""max	20.0	21.4	46.9	36.0	31.9	21.4	20.9	
100 av	17.1	13.8	51.0	37.4	24.0	18.8	18.4	3.8
""max	29.8	32.7	83.5	61.4	54.9	31.6	37.3	
			Process 8					
10 av	17	21	37	22	27	17	19	2.1
""max	31	38	55	43	49	21	27	
25 av	17	22	54	30	40	17	22	2.8
""max	33	35	74	58	69	30	37	
50 av	21	24	75	44	63	21	31	3.0
""max	44	41	89	79	93	34	62	
100 av	25	28	106	70	103	26	57	8.8
""max	62	53	120	113	135	36	101	

Table IV-2: Prediction Errors as Defined in Section (ii)

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 9					
10 av	28	21	32	19	28	16	25	3.8
""max	39	45	69	53	60	27	81	
25 av	41	36	52	34	51	28	50	5.5
""max	62	56	87	75	96	45	128	
50 av	51	46	64	46	76	36	72	9.0
""max	81	78	110	101	178	69	163	
100 av	58	57	91	76	118	55	93	14.8
""max	109	98	139	122	160	85	182	
			Process 10					
10 av	19	14	24	15	16	14	21	6.7
""max	29	114	78	53	135	28	80	
25 av	23	17	44	25	26	18	38	5.4
""max	30	120	90	76	162	26	133	
50 av	24	20	67	39	44	19	60	3.3
""max	33	123	106	95	180	28	166	
100 av	27	27	101	70	77	25	93	6.1
""max	52	127	131	124	190	47	183	

Table IV-2: Prediction Errors as Defined in Section (ii)

Timespan	ext1	ext2	reg+k	arma+k	reg	arma	armawt	disp
			Process 11					
10 av	21	20	21	21	20	21	39	5.1
""max	55	57	56	54	86	61	108	
25 av	24	23	33	30	29	27	70	4.7
""max	66	66	65	68	103	93	158	
50 av	26	26	52	46	39	35	102	9.1
""max	62	75	92	95	149	140	180	
100 av	36	35	91	84	65	52	143	8.7
""max	75	74	118	120	175	170	190	
			Process 12					
10 av	22	20	21	20	22	21	24	5.5
""max	49	34	31	31	33	33	75	
25 av	28	27	39	35	35	34	54	6.4
""max	44	47	53	60	80	58	134	
50 av	39	35	57	56	55	48	89	7.8
""max	61	66	94	85	118	74	169	
100 av	51	49	80	81	84	66	124	7.1
""max	80	75	128	122	116	111	185	

Table IV-2: Prediction Errors as Defined in Section (ii)

Process 1

	ext	reg+k	arma+k	reg	arma	armawt
Z1	1.019	1.016	1.016	1.018	1.018	1.015
Z2	1.026	1.011	1.010	1.019	1.019	1.020
Z3	1.031	1.029	1.031	1.030	1.031	1.031
Z4	1.034	1.002	1.002	1.014	1.014	1.015
Z5	1.023	1.001	1.000	1.012	1.012	1.016
Z6	1.024	1.021	1.021	1.023	1.023	1.021
Z7	1.023	1.016	1.016	1.021	1.021	1.022
Z8	1.029	1.016	1.015	1.022	1.022	1.024
Z9	1.032	1.011	1.014	1.020	1.022	1.023
Z10	1.025	1.026	1.026	1.026	1.026	1.024

Process 2

Z1	1.014	.896	.984	.995	1.024	.994
Z2	1.026	.967	1.007	1.020	1.022	1.015
Z3	1.025	1.001	1.033	1.020	1.031	1.006
Z4	1.020	.959	1.026	1.002	1.029	1.002
Z5	1.032	.957	1.028	.992	1.033	1.008
Z6	1.032	.959	1.001	.994	1.030	1.009
Z7	1.027	.940	.998	.994	1.015	.998
Z8	1.022	1.000	1.052	1.024	1.037	1.003
Z9	1.034	.950	1.004	.990	1.018	1.017
Z10	1.030	.968	1.027	1.005	1.028	1.012

Table IV-3: Estimates of Growth Factor, "c"

	ext	reg+k	arma+k	reg	arma	armawt
			Process 3			
Z1	1.006	.905	.998	.986	1.015	.876
Z2	1.029	.932	1.017	.986	1.031	.988
Z3	1.024	.926	.955	.967	.982	.883
Z4	1.030	.972	1.017	1.008	1.023	1.009
Z5	1.003	.760	.841	.959	.979	.783
Z6	1.032	.928	.978	.984	1.006	.987
Z7	1.027	.932	.988	.994	1.018	.998
Z8	1.014	.895	.987	.988	1.020	.889
Z9	1.025	.901	1.020	.989	1.022	.997
Z10	1.009	1.030	1.050	1.032	1.038	.964
			Process 4			
Z1	1.020	1.017	1.037	1.025	1.031	1.011
Z2	1.028	.766	1.015	.898	1.034	.890
Z3	1.022	.991	1.005	1.010	1.015	1.004
Z4	1.028	.968	1.030	1.004	1.031	.990
Z5	1.026	.929	1.002	.987	1.022	1.006
Z6	1.028	.980	.992	1.006	1.011	.997
Z7	1.033	.988	1.019	1.011	1.025	1.020
Z8	1.034	1.036	1.050	1.038	1.045	1.020
Z9	1.028	.978	1.011	1.010	1.024	1.013
Z10	1.023	1.029	1.042	1.030	1.034	.962

Table IV-3: Estimates of Growth Factor, "c"

	ext	reg+k	arma+k	reg	arma	armawt
			Process 5			
Z1	1.023	1.011	1.013	1.020	1.022	.987
Z2	1.021	.835	.905	.943	.984	.908
Z3	1.035	.993	1.011	1.014	1.021	1.022
Z4	1.026	.998	1.007	1.025	1.030	.982
Z5	1.020	.992	.994	1.014	1.015	.984
Z6	1.035	.926	.916	.983	.979	.909
Z7	1.020	1.015	1.018	1.023	1.024	1.015
Z8	1.028	1.013	1.013	1.022	1.022	1.008
Z9	1.029	1.005	1.008	1.018	1.020	.872
Z10	1.037	1.031	1.034	1.037	1.039	1.011
			Process 6			
Z1	1.023	1.016	1.015	1.021	1.021	1.010
Z2	1.020	1.002	1.013	1.024	1.027	.967
Z3	1.016	.935	.950	.999	1.006	.887
Z4	1.025	1.022	1.014	1.032	1.029	.986
Z5	1.045	.980	.982	1.002	1.003	.994
Z6	1.029	1.060	1.066	1.053	1.056	1.020
Z7	1.035	.890	.888	.952	.953	.918
Z8	1.053	1.050	1.047	1.053	1.052	1.052
Z9	1.004	.938	.951	.999	1.003	.903
Z10	1.030	1.042	1.043	1.039	1.040	.977

Table IV-3: Estimates of Growth Factor, "c"

	ext	reg+k	arma+k	reg	arma	armawt
			Process 7			
Z1	1.010	.989	.994	1.009	1.010	1.008
Z2	1.014	.998	1.002	1.011	1.012	1.012
Z3	1.016	1.002	1.011	1.014	1.016	1.016
Z4	1.018	.990	.997	1.009	1.011	1.013
Z5	1.012	.987	.993	1.007	1.009	1.009
Z6	1.013	1.001	1.009	1.011	1.012	1.011
Z7	1.012	1.002	1.004	1.012	1.012	1.012
Z8	1.015	.998	1.006	1.013	1.015	1.014
Z9	1.017	.988	1.004	1.010	1.014	1.015
Z10	1.013	1.005	1.008	1.014	1.013	1.012
			Process 8			
Z1	1.008	.673	.740	.980	1.011	.996
Z2	1.013	.906	.969	1.002	1.011	1.007
Z3	1.014	.933	1.017	1.004	1.016	.999
Z4	1.010	.834	.914	.988	1.014	1.002
Z5	1.016	.884	.979	.985	1.018	1.013
Z6	1.017	.897	1.012	.987	1.016	1.004
Z7	1.014	.847	.992	.986	1.010	1.004
Z8	1.012	.844	.981	.994	1.014	1.002
Z9	1.018	.889	.957	.985	1.015	1.006
Z10	1.016	.883	.942	.993	1.014	1.009

Table IV-3: Estimates of Growth Factor, "c"

	ext	reg+k	arma+k	reg	arma	armawt
			Process 9			
Z1	1.008	.786	1.003	.978	1.012	.966
Z2	1.016	.834	1.014	.978	1.017	1.009
Z3	1.015	.839	.961	.967	1.001	.972
Z4	1.016	.883	.951	.994	1.014	1.005
Z5	1.004	.475	.619	.966	.999	.932
Z6	1.018	.863	.958	.981	1.009	1.000
Z7	1.015	.835	1.000	.987	1.012	1.006
Z8	1.011	.734	.886	.973	1.012	.924
Z9	1.013	.743	.833	.979	1.014	1.009
Z10	1.006	.892	.973	.998	1.016	.988
			Process 10			
Z1	1.011	.958	.983	1.008	1.014	1.006
Z2	1.015	.662	.791	.929	1.015	.914
Z3	1.011	.936	.973	1.002	1.010	1.002
Z4	1.015	.867	.940	.988	1.016	.987
Z5	1.014	.851	.987	.986	1.014	1.011
Z6	1.015	.937	.987	1.000	1.009	.985
Z7	1.018	.925	.986	.997	1.017	.978
Z8	1.018	.971	1.005	1.008	1.021	1.007
Z9	1.014	.903	1.001	.998	1.014	1.003
Z10	1.012	.958	.985	1.008	1.015	.950

Table IV-3: Estimates of Growth Factor, "c"

	ext	reg+k	arma+k	reg	arma	armawt
			Process 11			
Z1	1.013	.968	.975	1.007	1.009	.972
Z2	1.012	.731	.849	.944	.985	.888
Z3	1.023	.967	.988	1.003	1.012	.973
Z4	1.014	.916	.935	1.004	1.011	.957
Z5	1.008	.870	.879	.990	.993	.953
Z6	1.020	.831	.791	.959	.950	.828
Z7	1.012	.964	.978	1.009	1.011	.998
Z8	1.015	.968	.991	1.010	1.012	.994
Z9	1.017	.971	.976	1.006	1.008	.839
Z10	1.019	.996	.999	1.015	1.017	.997
			Process 12			
Z1	1.013	.980	.987	1.008	1.010	.997
Z2	1.012	.938	.978	1.008	1.016	.966
Z3	1.010	.839	.874	.988	1.000	.867
Z4	1.013	.943	.962	1.006	1.015	.973
Z5	1.024	.946	.971	.993	1.002	.981
Z6	1.015	1.018	1.028	1.023	1.025	1.001
Z7	1.020	.834	.891	.964	.990	.941
Z8	1.029	1.017	1.027	1.027	1.032	1.012
Z9	1.005	.764	.822	.993	1.008	.971
Z10	1.015	.995	1.007	1.013	1.018	.949

Table IV-3: Estimates of Growth Factor, "c"

Process 1 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	11	14	10	29	22	6	9	5	25	5
ext2	12	8	12	12	8	6	6	6	7	5
reg+k	11	10	12	23	14	6	4	10	6	4
arma+k	10	8	12	23	15	5	4	11	4	4
reg	11	8	12	22	13	6	4	9	5	4
arma	11	7	12	22	13	5	5	9	4	4
armawt	9	8	12	21	11	5	5	8	4	4
Predicting Z(101)-Z(125) from Z(100)										
ext1	11	23	15	33	16	9	11	9	24	5
ext2	13	11	13	11	19	9	8	14	8	5
reg+k	11	10	12	38	38	11	5	27	21	4
arma+k	10	9	13	38	39	11	5	28	17	5
reg	12	7	13	33	32	10	6	23	15	4
arma	12	6	13	34	34	11	7	23	12	5
armawt	8	6	13	32	28	12	7	21	11	5
Predicting Z(101)-Z(150) from Z(100)										
ext1	24	19	24	36	23	8	27	30	39	23
ext2	25	10	22	8	20	8	24	35	22	23
reg+k	20	37	19	66	49	9	11	60	30	22
arma+k	20	36	21	66	50	10	12	61	23	21
reg	24	25	20	57	38	8	18	51	18	22
arma	24	24	21	57	40	9	20	52	13	20
armawt	16	21	21	53	30	11	21	48	12	26

Table IV-4: Errors in Prediction & Miscellany

Process 1 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	23	20	50	53	26	13	24	54	35	44
ext2	24	26	47	24	19	13	21	59	19	44
reg+k	27	84	40	104	83	22	33	102	80	43
arma+k	27	84	46	104	86	22	32	104	70	41
reg	24	63	43	87	61	16	21	87	55	42
arma	24	62	46	87	64	17	21	89	46	39
armawt	31	55	45	80	45	25	20	82	39	36

Predicting Z(1)-Z(100) from Z(1)

ext1	12	8	12	16	11	9	8	6	11	13
ext2	12	10	39	20	11	16	25	9	20	17
reg+k	16	27	21	51	30	23	14	39	42	14
arma+k	15	30	35	51	31	20	16	40	38	13
reg	12	47	42	103	56	12	37	31	76	16
arma	12	47	39	104	59	12	35	33	68	15
armawt	22	39	41	98	39	12	33	25	61	23

Iterations and Significance of ARMA vs. regression

its/ak	0	1	1	0	3	3	3	1	3	1
its/a	0	1	3	0	3	3	3	1	3	1
its/aw	3	2	1	8	6	4	2	1	6	4
p/ak	.49	.00	.05	.50	.25	.01	.15	.42	.02	.26
p/a	.49	.00	.05	.50	.23	.01	.14	.43	.02	.26

Table IV-4: Errors in Prediction & Miscellany

Process 2 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	48	15	8	26	24	22	29	27	18	11
ext2	18	12	22	16	25	37	23	37	18	14
reg+k	40	16	15	28	45	58	44	30	40	33
arma+k	22	17	22	18	26	36	22	35	22	11
reg	20	12	19	20	42	53	39	36	34	25
arma	18	13	21	19	28	19	18	31	19	11
armawt	33	31	11	15	23	33	28	21	15	20
Predicting Z(101)-Z(125) from Z(100)										
ext1	55	14	21	34	26	22	26	22	15	19
ext2	21	12	37	22	19	37	24	44	19	17
reg+k	76	25	19	62	66	88	73	28	78	53
arma+k	37	22	43	14	27	53	38	61	43	16
reg	39	12	31	43	61	80	65	45	71	35
arma	18	13	41	15	32	20	27	48	33	16
armawt	55	37	14	36	23	50	50	20	28	26
Predicting Z(101)-Z(150) from Z(100)										
ext1	76	21	20	55	34	25	25	28	22	33
ext2	41	19	36	44	20	30	29	60	18	27
reg+k	116	53	28	105	98	111	107	27	112	80
arma+k	73	42	51	20	32	63	66	102	62	26
reg	81	23	27	83	94	103	101	63	106	50
arma	24	23	47	18	43	23	47	80	43	27
armawt	97	52	34	76	44	60	86	24	39	32

Table IV-4: Errors in Prediction & Miscellany

Process 2 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	104	21	23	90	32	24	30	26	50	31
ext2	73	20	30	81	20	25	30	60	39	25
reg+k	149	96	78	148	141	147	142	50	143	131
arma+k	118	76	58	40	28	107	105	138	91	22
reg	129	37	31	131	141	144	140	66	144	104
arma	33	30	48	32	43	22	74	105	55	24
armawt	140	76	86	127	96	103	127	74	52	110

Predicting Z(1)-Z(100) from Z(1)

ext1	26	18	18	22	18	20	25	21	18	21
ext2	26	25	19	28	18	21	33	32	18	47
reg+k	47	58	62	54	75	81	61	80	85	67
arma+k	66	50	24	67	44	78	36	180	67	22
reg	91	52	33	76	142	137	142	38	148	134
arma	60	39	33	56	19	21	79	83	75	38
armawt	94	64	95	85	108	101	132	86	78	116

Iterations and Significance of ARMA vs. regression

its/ak	10	3	8	10	10	10	10	8	9	10
its/a	4	1	4	4	10	10	6	6	4	9
its/aw	7	6	7	4	8	9	9	4	5	7
p/ak	.00	.20	.00	.00	.00	.01	.00	.00	.00	.00
p/a	.00	.17	.00	.00	.00	.00	.00	.00	.00	.00

Table IV-4: Errors in Prediction & Miscellany

Process 3 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	90	16	84	34	34	65	48	83	16	93
ext2	17	44	26	30	21	16	38	51	14	23
reg+k	43	66	26	14	18	36	14	77	42	28
arma+k	17	17	29	18	16	17	22	52	23	35
reg	26	62	50	18	42	38	20	62	23	29
arma	12	15	46	20	23	14	27	40	20	31
armawt	84	36	98	14	120	27	18	106	34	29

Predicting Z(101)-Z(125) from Z(100)

ext1	107	34	108	98	48	67	91	96	44	109
ext2	43	61	55	96	40	35	83	67	45	28
reg+k	87	106	48	58	40	73	40	112	69	26
arma+k	43	43	49	81	39	55	58	77	43	42
reg	66	103	67	74	73	80	52	93	47	27
arma	31	29	58	84	47	50	69	53	43	33
armawt	115	83	141	70	167	70	50	153	52	80

Predicting Z(101)-Z(150) from Z(100)

ext1	137	52	139	120	49	52	139	118	95	122
ext2	92	71	86	118	36	53	134	94	98	42
reg+k	132	132	56	52	40	119	75	143	52	30
arma+k	93	62	55	94	40	103	104	114	83	69
reg	121	134	82	77	110	129	92	131	41	33
arma	73	49	58	101	67	95	121	77	87	48
armawt	173	117	168	74	184	121	92	177	46	126

Table IV-4: Errors in Prediction & Miscellany

Process 3 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	165	77	132	122	68	43	164	127	120	142
ext2	141	70	75	119	85	77	161	105	124	79
reg+k	165	142	88	85	77	157	68	163	70	49
arma+k	143	51	88	77	77	148	111	139	102	109
reg	161	153	143	64	157	165	84	159	68	54
arma	124	83	126	90	134	141	146	79	109	76
armawt	187	136	185	61	192	161	90	189	58	142

Predicting Z(1)-Z(100) from Z(1)

ext1	60	20	43	21	43	31	22	42	18	42
ext2	69	64	95	25	62	43	22	43	30	60
reg+k	68	67	78	71	43	74	67	54	62	129
arma+k	85	34	75	32	43	59	59	54	71	148
reg	85	165	137	113	135	162	135	120	145	130
arma	93	26	115	45	90	132	53	45	24	143
armawt	184	160	178	108	190	160	127	185	129	132

Iterations and Significance of ARMA vs. regression

its/ak	9	10	4	5	7	9	10	10	10	3
its/a	2	9	2	6	3	3	9	4	7	2
its/aw	3	9	6	9	5	8	9	9	10	4
p/ak	.00	.00	.18	.00	.24	.00	.00	.00	.00	.15
p/a	.00	.00	.12	.00	.03	.00	.00	.00	.00	.16

Table IV-4: Errors in Prediction & Miscellany

Process 4 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	34	30	8	29	26	42	42	38	29	32
ext2	7	113	9	31	5	15	29	22	29	10
reg+k	7	110	19	24	30	27	11	20	35	9
arma+k	6	25	15	31	9	18	23	13	33	11
reg	5	145	15	23	24	25	16	19	30	9
arma	5	26	13	32	19	16	25	14	29	9
armawt	12	100	18	27	7	23	18	31	36	40

Predicting Z(101)-Z(125) from Z(100)

ext1	43	35	12	23	30	53	56	42	25	38
ext2	18	119	13	26	13	15	42	31	25	17
reg+k	17	129	37	41	61	40	14	28	45	16
arma+k	10	35	28	29	14	29	27	25	31	24
reg	12	172	28	26	54	34	15	28	30	17
arma	9	26	22	31	21	22	32	25	24	19
armawt	31	150	36	48	20	39	21	46	37	87

Predicting Z(101)-Z(150) from Z(100)

ext1	60	39	17	31	25	50	63	44	33	38
ext2	35	125	19	35	15	17	49	31	31	21
reg+k	36	149	65	73	100	77	47	26	64	25
arma+k	11	50	47	39	40	63	24	26	28	48
reg	23	186	47	46	100	66	26	23	31	26
arma	9	20	35	43	18	49	30	20	22	34
armawt	60	176	62	82	54	79	19	65	36	130

Table IV-4: Errors in Prediction & Miscellany

Process 4 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	90	49	35	25	30	38	91	38	40	55
ext2	70	132	36	27	43	34	79	35	37	33
reg+k	74	170	113	130	145	129	87	41	113	22
arma+k	17	83	90	30	100	119	22	87	55	72
reg	50	193	88	107	149	118	43	47	66	24
arma	19	17	67	34	46	100	40	72	19	41
armawt	105	188	108	138	112	132	17	68	64	161
Predicting Z(1)-Z(100) from Z(1)										
ext1	16	29	23	18	16	23	18	16	15	17
ext2	16	36	29	28	22	45	19	18	22	18
reg+k	54	76	52	74	66	41	73	57	63	53
arma+k	25	55	47	18	57	33	42	190	31	76
reg	29	186	52	98	135	127	105	22	99	47
arma	58	37	32	23	19	112	50	49	29	66
armawt	53	186	74	130	89	142	70	78	81	162
Iterations and Significance of ARMA vs. regression										
its/ak	10	9	7	10	9	7	8	5	10	9
its/a	3	10	3	10	9	2	5	6	7	1
its/aw	4	6	3	3	10	9	9	6	10	6
p/ak	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
p/a	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table IV-4: Errors in Prediction & Miscellany

Process 5 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	28	33	60	19	41	33	20	21	13	29
ext2	19	57	33	23	23	21	6	17	16	15
reg+k	21	78	17	15	29	48	5	11	14	14
arma+k	20	71	19	16	29	50	6	11	14	15
reg	19	89	21	22	24	42	7	13	14	15
arma	18	69	22	23	24	43	7	13	14	15
armawt	35	102	22	24	40	78	4	6	83	16
Predicting Z(101)-Z(125) from Z(100)										
ext1	21	61	56	23	36	36	22	19	20	29
ext2	18	65	30	22	19	20	21	16	27	40
reg+k	18	83	29	21	33	79	19	15	13	35
arma+k	17	77	20	17	32	82	20	15	13	38
reg	17	114	21	21	22	72	22	13	16	39
arma	18	78	18	25	22	76	23	13	16	42
armawt	48	138	18	61	59	133	18	25	141	18
Predicting Z(101)-Z(150) from Z(100)										
ext1	17	64	75	32	51	29	23	21	22	29
ext2	16	82	52	20	33	25	20	19	30	44
reg+k	21	124	45	50	70	13	19	34	31	35
arma+k	17	120	26	36	68	125	20	34	29	42
reg	14	156	27	20	45	122	24	23	19	45
arma	15	128	24	25	42	125	26	23	18	50
armawt	86	170	26	109	107	167	21	54	172	36

Table IV-4: Errors in Prediction & Miscellany

Process 5 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	48	63	85	55	84	46	66	39	27	32
ext2	43	87	62	32	68	31	54	41	36	50
reg+k	71	153	100	105	122	154	67	90	71	33
arma+k	65	150	59	86	119	155	56	91	65	45
reg	49	179	61	34	89	159	43	68	36	51
arma	46	161	33	26	84	161	39	68	31	60
armawt	139	185	27	154	152	184	75	115	186	106
Predicting Z(1)-Z(100) from Z(1)										
ext1	28	36	29	24	35	46	27	22	37	20
ext2	28	43	76	25	48	47	54	41	52	20
reg+k	54	63	74	79	59	83	42	57	79	97
arma+k	53	62	58	80	59	83	38	57	79	92
reg	30	176	68	26	67	159	43	29	49	20
arma	28	145	50	30	62	162	40	29	46	24
armawt	135	184	75	151	149	184	76	81	185	111
Iterations and Significance of ARMA vs. regression										
its/ak	2	8	8	1	1	1	2	0	1	1
its/a	2	3	4	1	1	1	2	0	1	1
its/aw	8	9	10	10	9	10	7	6	9	6
p/ak	.40	.03	.02	.41	.45	.44	.20	.50	.42	.38
p/a	.39	.01	.00	.37	.43	.47	.18	.50	.41	.37

Table IV-4: Errors in Prediction & Miscellany

Process 6 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	29	42	16	26	50	23	109	22	36	35
ext2	26	21	11	28	7	30	15	17	40	43
reg+k	26	15	16	27	29	43	21	16	27	45
arma+k	26	17	12	23	28	47	21	14	28	46
reg	25	22	5	30	26	41	56	17	35	45
arma	25	22	7	28	26	43	55	16	37	45
armawt	30	18	67	12	31	23	73	16	41	22

Predicting Z(101)-Z(125) from Z(100)

ext1	42	36	35	46	58	28	104	28	34	35
ext2	36	25	31	30	17	55	22	30	44	53
reg+k	41	15	56	29	55	85	52	28	26	63
arma+k	42	20	51	30	53	90	52	27	27	65
reg	38	29	38	31	50	80	112	30	38	62
arma	38	30	33	29	48	83	111	29	42	62
armawt	52	58	129	58	59	43	133	29	101	37

Predicting Z(101)-Z(150) from Z(100)

ext1	57	58	69	81	75	35	119	40	36	63
ext2	50	60	56	52	40	64	33	28	41	81
reg+k	60	56	109	51	89	116	67	28	36	98
arma+k	61	56	105	65	88	123	67	29	34	100
reg	53	63	88	40	82	107	148	28	34	96
arma	54	65	75	43	77	112	147	28	39	96
armawt	78	101	166	118	97	46	165	28	149	69

Table IV-4: Errors in Prediction & Miscellany

Process 6 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	68	48	66	102	113	92	140	52	92	102
ext2	63	47	51	73	86	114	66	65	69	116
reg+k	85	63	135	74	126	157	101	56	100	139
arma+k	87	45	131	99	125	162	101	47	98	141
reg	70	58	112	42	118	152	174	68	83	135
arma	71	67	87	55	114	155	174	61	72	136
armawt	112	151	184	157	137	86	183	62	176	159
Predicting Z(1)-Z(100) from Z(1)										
ext1	16	32	33	22	31	18	48	40	42	25
ext2	31	54	38	22	57	18	54	46	50	74
reg+k	45	85	51	90	101	190	88	150	50	96
arma+k	46	90	50	87	100	190	88	151	51	90
reg	25	68	99	49	128	108	175	47	45	108
arma	23	80	74	35	126	116	175	44	46	109
armawt	52	146	185	138	139	50	182	44	178	140
Iterations and Significance of ARMA vs. regression										
its/ak	1	1	2	1	0	1	0	2	2	2
its/a	1	1	1	1	0	1	0	2	2	2
its/aw	8	5	8	10	4	9	9	4	3	10
p/ak	.40	.37	.39	.43	.49	.32	.50	.15	.40	.04
p/a	.40	.35	.29	.45	.49	.36	.50	.15	.31	.05

Table IV-4: Errors in Prediction & Miscellany

Process 7 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	7	9	6	15	11	5	6	4	13	4
ext2	10	6	8	9	7	6	7	7	4	6
reg+k	6	4	11	16	12	6	6	11	8	4
arma+k	4	5	9	13	9	6	4	6	3	4
reg	9	5	9	13	9	5	7	8	5	5
arma	6	5	7	10	7	5	5	4	5	4
armawt	5	5	7	8	7	6	5	4	5	4

Predicting Z(101)-Z(125) from Z(100)

ext1	7	12	8	17	9	6	7	6	13	4
ext2	11	8	7	9	12	6	9	10	6	6
reg+k	6	6	10	26	26	12	5	23	22	5
arma+k	6	6	6	20	21	9	5	14	8	6
reg	9	5	7	19	19	7	8	15	11	6
arma	6	5	7	14	15	7	6	6	5	4
armawt	5	5	8	10	14	8	6	8	5	5

Predicting Z(101)-Z(150) from Z(100)

ext1	13	10	13	19	12	5	14	17	20	13
ext2	17	7	11	7	12	5	15	21	10	11
reg+k	10	23	16	44	35	18	6	47	35	21
arma+k	8	20	7	36	27	9	5	33	12	20
reg	14	10	8	29	21	7	15	32	14	10
arma	12	9	11	21	15	6	12	18	8	13
armawt	7	8	12	15	14	8	11	21	8	17

Table IV-4: Errors in Prediction & Miscellany

Process 7 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	13	11	26	28	13	8	12	30	18	24
ext2	14	13	23	12	11	8	13	33	9	21
reg+k	48	58	27	74	62	41	28	84	76	46
arma+k	42	50	7	61	51	20	25	61	40	41
reg	15	27	13	44	34	15	13	55	37	20
arma	13	25	22	30	22	10	11	32	11	24
armawt	20	23	24	18	22	16	11	37	10	23

Predicting Z(1)-Z(100) from Z(1)

ext1	7	5	7	9	6	5	5	5	7	7
ext2	9	6	26	10	7	12	14	7	9	11
reg+k	16	14	12	23	17	19	10	20	23	13
arma+k	13	10	6	20	17	11	6	18	17	7
reg	8	19	35	53	26	8	15	19	44	11
arma	7	16	20	41	16	7	14	5	21	10
armawt	12	14	22	30	16	7	14	7	19	17

Iterations and Significance of ARMA vs. regression

its/ak	9	7	9	7	9	9	7	10	9	8
its/a	3	4	4	3	2	4	3	2	3	3
its/aw	3	2	5	6	3	3	2	5	4	2
p/ak	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
p/a	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table IV-4: Errors in Prediction & Miscellany

Process 8 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	31	16	9	19	21	21	18	23	18	11
ext2	18	11	22	18	30	38	25	33	19	16
reg+k	47	27	16	38	48	55	42	32	39	40
arma+k	43	24	18	23	22	20	16	21	26	27
reg	23	11	17	24	43	49	39	26	30	26
arma	18	20	18	17	21	18	14	21	15	13
armawt	27	25	11	15	19	25	18	22	16	17
Predicting Z(101)-Z(125) from Z(100)										
ext1	33	14	17	21	19	18	17	19	15	14
ext2	19	14	32	19	25	35	25	33	19	16
reg+k	62	47	25	61	62	74	58	49	66	59
arma+k	58	33	31	46	24	19	19	18	45	39
reg	46	12	21	46	61	69	60	20	58	39
arma	19	19	30	14	22	16	15	21	14	14
armawt	37	27	14	21	16	31	24	21	24	19
Predicting Z(101)-Z(150) from Z(100)										
ext1	44	19	17	32	23	20	18	20	18	21
ext2	25	18	32	30	22	29	27	41	16	18
reg+k	83	74	52	86	80	89	78	60	87	75
arma+k	79	55	35	75	39	19	32	20	64	54
reg	87	29	19	82	91	93	92	24	89	58
arma	23	25	34	16	28	19	21	31	14	16
armawt	62	38	27	43	18	37	39	22	34	19

Table IV-4: Errors in Prediction & Miscellany

Process 9 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	39	16	29	19	22	34	27	39	16	47
ext2	12	45	15	25	15	18	24	40	15	27
reg+k	44	59	28	21	13	34	22	69	44	26
arma+k	13	15	23	15	13	16	19	53	37	20
reg	23	60	34	15	31	34	14	58	17	23
arma	11	15	20	15	15	9	21	24	14	27
armawt	38	18	34	14	52	12	17	81	19	20
Predicting Z(101)-Z(125) from Z(100)										
ext1	50	24	51	50	31	37	54	46	30	62
ext2	21	56	34	56	26	29	51	47	32	24
reg+k	72	85	38	31	25	59	34	87	62	62
arma+k	25	24	32	27	25	43	41	75	54	37
reg	57	96	57	32	58	68	28	89	41	28
arma	17	21	30	44	25	27	45	31	30	23
armawt	78	35	53	33	98	37	37	128	31	44
Predicting Z(101)-Z(150) from Z(100)										
ext1	73	31	68	59	26	31	81	60	42	70
ext2	46	60	45	66	26	39	78	60	50	24
reg+k	103	103	36	41	28	91	30	110	56	86
arma+k	54	31	30	27	28	77	60	101	47	58
reg	105	127	91	39	102	112	27	178	54	41
arma	34	28	27	49	27	47	69	43	44	22
armawt	126	45	83	29	143	69	52	163	32	72

Table IV-4: Errors in Prediction & Miscellany

Process 9 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	109	33	61	58	42	28	101	62	53	86
ext2	87	45	40	65	55	52	98	62	63	40
reg+k	139	114	83	81	67	128	32	128	71	114
arma+k	100	31	78	66	67	119	65	122	63	94
reg	151	155	145	72	151	153	58	160	105	75
arma	68	36	69	43	67	79	85	43	58	31
armawt	164	39	139	37	173	112	54	182	35	86
Predicting Z(1)-Z(100) from Z(1)										
ext1	33	18	28	18	26	22	18	25	16	27
ext2	44	40	49	18	33	24	25	27	18	31
reg+k	40	42	50	45	28	47	41	37	37	36
arma+k	36	18	44	44	28	36	29	36	38	39
reg	108	153	141	107	134	143	129	137	136	40
arma	56	18	53	21	30	60	24	26	16	72
armawt	134	55	145	68	165	99	61	175	30	79
Iterations and Significance of ARMA vs. regression										
its/ak	10	10	10	10	9	10	10	10	10	10
its/a	5	10	5	7	6	8	10	6	10	3
its/aw	6	10	9	9	10	6	10	2	10	4
p/ak	.00	.00	.00	.00	.09	.00	.00	.00	.00	.00
p/a	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table IV-4: Errors in Prediction & Miscellany

Process 10 : Predicting Z(101)-Z(110) from Z(100)										
method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	18	28	6	25	15	21	23	21	29	18
ext2	6	114	4	27	6	16	14	17	29	10
reg+k	15	78	18	32	28	26	14	30	37	17
arma+k	10	53	14	27	7	11	9	16	29	12
reg	7	135	8	22	18	22	7	22	28	10
arma	6	26	5	25	16	8	19	11	28	10
armawt	9	80	11	29	9	25	19	21	33	39
Predicting Z(101)-Z(125) from Z(100)										
ext1	24	30	8	19	18	27	30	30	22	25
ext2	12	120	6	22	12	15	20	28	23	17
reg+k	38	90	38	54	50	38	33	51	52	39
arma+k	25	76	27	41	20	19	10	30	24	25
reg	15	162	17	32	42	31	15	35	27	19
arma	10	26	9	21	19	13	26	24	22	16
armawt	19	133	20	44	12	46	44	36	33	87
Predicting Z(101)-Z(150) from Z(100)										
ext1	32	28	13	25	15	26	33	28	21	27
ext2	19	123	11	29	11	17	23	25	23	21
reg+k	66	106	62	73	72	61	62	78	67	59
arma+k	47	95	47	60	44	37	25	37	19	37
reg	27	180	33	58	78	52	42	46	35	25
arma	12	21	15	27	16	22	28	18	19	21
armawt	33	166	36	71	12	83	88	48	35	131

Table IV-4: Errors in Prediction & Miscellany

Process 10 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	52	31	22	19	18	20	51	23	23	32
ext2	40	127	19	22	25	26	42	22	25	23
reg+k	106	131	97	110	111	100	89	105	97	97
arma+k	88	124	83	102	92	78	44	41	31	72
reg	55	190	66	113	129	92	73	54	71	42
arma	25	20	28	21	18	47	43	30	19	18
armawt	63	183	69	122	34	132	134	58	58	163

Predicting Z(1)-Z(100) from Z(1)

ext1	11	28	16	17	13	17	18	14	15	15
ext2	12	31	18	22	17	24	18	14	15	15
reg+k	27	47	33	43	38	30	44	40	38	31
arma+k	30	49	30	40	28	16	43	40	27	36
reg	19	177	48	110	115	88	101	52	88	29
arma	21	28	17	18	14	49	18	19	16	23
armawt	26	182	45	111	19	133	143	59	59	165

Iterations and Significance of ARMA vs. regression

its/ak	10	10	10	10	10	9	10	10	10	10
its/a	5	10	4	10	10	3	10	9	10	9
its/aw	4	10	6	9	10	4	9	3	10	3
p/ak	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
p/a	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table IV-4: Errors in Prediction & Miscellany

Process 11 : Predicting Z(101)-Z(110) from Z(100)										
method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	22	28	55	14	28	32	10	18	15	21
ext2	18	57	29	26	21	23	6	23	18	14
reg+k	24	56	13	18	36	45	8	14	18	12
arma+k	23	54	17	16	36	48	6	14	17	11
reg	19	86	18	20	28	46	5	19	16	12
arma	18	61	20	21	26	50	5	19	15	13
armawt	35	108	13	29	47	103	6	9	93	13
Predicting Z(101)-Z(125) from Z(100)										
ext1	18	66	47	16	26	31	17	17	19	28
ext2	18	66	23	26	18	19	19	21	27	38
reg+k	26	63	35	45	48	65	24	16	18	22
arma+k	23	63	24	38	47	68	21	15	17	24
reg	17	103	24	18	38	84	17	17	16	33
arma	17	67	19	20	34	93	18	17	17	35
armawt	53	144	63	73	84	158	20	23	150	17
Predicting Z(101)-Z(150) from Z(100)										
ext1	14	55	62	20	36	27	17	17	22	23
ext2	16	75	40	24	27	18	22	21	32	35
reg+k	43	89	50	75	72	92	39	35	38	23
arma+k	37	87	32	69	72	95	31	32	35	21
reg	16	149	31	24	72	131	18	18	17	27
arma	13	109	24	18	64	140	20	17	16	31
armawt	95	172	99	122	131	180	31	51	176	40

Table IV-4: Errors in Prediction & Miscellany

Process 11 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	34	44	75	32	64	43	37	23	27	23
ext2	32	74	53	26	55	27	31	22	37	35
reg+k	88	111	90	112	110	118	91	84	77	56
arma+k	82	109	62	108	110	120	83	78	7	48
reg	49	175	59	55	122	165	41	41	34	23
arma	40	144	25	29	112	170	33	35	24	29
armawt	143	187	147	162	167	190	85	106	188	98
Predicting Z(1)-Z(100) from Z(1)										
ext1	23	37	28	25	33	43	21	20	33	16
ext2	23	44	74	25	43	47	41	33	44	16
reg+k	36	47	54	44	40	60	28	37	53	48
arma+k	36	46	51	43	40	61	27	36	52	46
reg	40	172	60	60	111	165	53	25	48	25
arma	30	126	41	31	100	171	43	25	41	18
armawt	143	186	160	162	168	188	97	87	188	103
Iterations and Significance of ARMA vs. regression										
its/ak	2	8	9	3	0	4	5	2	2	1
its/a	1	4	6	1	2	1	2	2	2	1
its/aw	2	3	10	4	4	5	4	3	4	5
p/ak	.28	.05	.00	.38	.48	.24	.11	.23	.39	.43
p/a	.22	.00	.00	.23	.39	.39	.08	.19	.34	.40

Table IV-4: Errors in Prediction & Miscellany

Process 12 : Predicting Z(101)-Z(110) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	22	36	10	15	30	20	49	9	27	29
ext2	22	23	12	29	12	23	13	19	29	34
reg+k	26	6	22	13	31	24	25	14	31	28
arma+k	25	8	18	16	24	28	19	19	27	31
reg	23	20	6	24	29	26	33	17	25	32
arma	22	18	7	30	20	28	18	20	27	33
armawt	28	18	75	13	32	14	49	9	31	25
Predicting Z(101)-Z(125) from Z(100)										
ext1	27	29	28	33	33	25	44	19	28	26
ext2	26	30	31	30	13	47	14	28	36	37
reg+k	44	27	45	47	53	52	52	22	36	27
arma+k	41	11	42	40	41	60	44	28	31	33
reg	32	25	34	30	49	55	80	26	28	35
arma	30	27	29	31	33	58	48	31	35	39
armawt	47	50	134	57	61	28	105	21	48	68
Predicting Z(101)-Z(150) from Z(100)										
ext1	35	54	42	59	46	32	61	22	26	33
ext2	33	66	35	40	28	53	28	25	28	48
reg+k	66	62	83	94	80	62	62	25	49	27
arma+k	61	54	81	85	68	77	54	25	43	39
reg	44	60	78	55	80	69	118	23	37	45
arma	41	66	52	34	56	74	67	30	29	53
armawt	71	89	169	114	99	32	145	41	93	118

Table IV-4: Errors in Prediction & Miscellany

Process 12 : Predicting Z(101)-Z(200) from Z(100)

method	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
ext1	42	40	37	74	71	49	80	38	61	48
ext2	40	58	29	52	49	75	45	51	50	64
reg+k	101	86	98	128	108	92	86	27	92	33
arma+k	94	64	97	122	96	116	80	47	87	46
reg	64	46	111	82	116	104	157	43	90	59
arma	57	65	63	40	79	111	107	64	44	75
armawt	108	142	185	155	141	29	173	55	144	168
Predicting Z(1)-Z(100) from Z(1)										
ext1	14	25	29	20	23	16	33	30	22	24
ext2	26	35	30	20	33	16	39	34	25	34
reg+k	30	45	39	40	60	52	58	82	27	47
arma+k	31	51	38	39	55	48	57	75	29	50
reg	20	28	106	43	117	44	165	31	59	30
arma	18	48	62	24	89	54	134	43	31	43
armawt	69	144	186	143	139	77	173	77	128	164
Iterations and Significance of ARMA vs. regression										
its/ak	2	7	7	7	5	1	9	3	10	5
its/a	2	2	2	2	3	1	3	2	4	2
its/aw	3	2	3	2	3	4	3	10	10	3
p/ak	.14	.07	.29	.26	.08	.37	.02	.03	.00	.07
p/a	.11	.03	.08	.11	.06	.37	.00	.03	.00	.05

Table IV-4: Errors in Prediction & Miscellany

(VI) NATIONALISM AND SOCIAL COMMUNICATIONS: A TEST CASE FOR MATHEMATICAL APPROACHES

(i) INTRODUCTION AND SUMMARY

In the previous chapter, we have emphasized the importance of carrying out statistical research within the context of a broader analytic effort. The substantive goal of this thesis, however, in political science, was to carry through an analytic point of view, already developed by Karl Deutsch, and formulated mathematically with the assistance of Robert Solow(1). In the first phase of this research, carried out in 1971, we attempted to develop the original Deutsch-Solow model as a predictive model of national assimilation and political mobilization; more precisely, we attempted to predict such indicators of national assimilation as language or ethnicity (see Table VI-24), and such indicators of social mobilization as urbanization or literacy (see Table VI-23). (Note that these indicators were suggested originally by Karl Deutsch, not as operational definitions of nationalism, but as usable series of numerical data with some sort of correlation, albeit noisy, with the underlying concepts he has discussed.)

Even though the Deutsch-Solow model seems very simple, from the mathematician's point of view, the existing statistical routines turned out to be unable to cope effectively with even this level of complexity. Because of this result, which we found rather surprising at first, we found it necessary to abandon, in this context, the more ambitious goal of predicting long-term political trends by way of more interesting, complex models. We have, instead, developed two distinct strands of thought - one methodological and the other substantive - which may be prerequisites to success in the more ambitious undertakings of the future. Even in their present form, however, these two strands do offer predictions and insights, respectively, of some relevance to the decision-maker concerned with nationalism.

First of all, we have developed a new methodology for statistical analysis, the "robust method" of section (xi) of Chapter (II), able to deal effectively with prediction over time. This method emerged from the study of the Deutsh-Solow model and of similar simple models. In sections (ii) and (iii) of this chapter, we will describe the empirical results, based on the Deutsch-Kravitz data from more than a dozen nations, which led us to this new approach. These results

LEGEND TO FIGURES VI-1 THROUGH VI-4

The figures on the next four pages describe the average percentage errors which we found when making long-term predictions of our four variables - the sizes of the mobilized, underlying, assimilated and differentiated populations - in a variety of different cases, by use of different estimation techniques. In each "case" (i.e. a nation and a choice of data to study in that nation), we calculate the root-mean-square ("RMS") average of the errors in the predictions made to all different years for which data were available; this may be thought of as taking an average across different intervals of prediction. For each of our three basic techniques - regression, ARMA and the robust method (GRR or EXTRAP) - and for each variable, we have drawn a curve which represents the distribution of average error size from case to case. These distribution curves are like the distribution curves for college board scores; to find out how bad the errors were for the 20th percentile down from the top, we look at 20% on the horizontal axis, and then look up at our curves to see how high the prediction errors go, in the vertical direction. Notice that the vertical axis is spread out at the bottom, and compressed at the top, to allow us to fit the whole curve on one page; it is still correct, however, when a curve is exactly halfway between, say, 50% and 70%, to conclude that the error was exactly 60%. Thus in comparing the area under different curves, it is important to note that the horizontal line at "5%" should be thought of as the base of the graph, in regions where the error percentages are between 10% and 50%, in order to compensate for the spreading out at the bottom.

The distributions for regression, ARMA and GRR were drawn from Tables VI-15 through VI-20, from the columns labelled "Uni." and "ext1". They all represent predictions based on the reduced form of the Deutsch-Solow model, equations (6.1) and (6.2) with the "bD" and "fU" terms removed, for the same cases, defined in Tables VI-23 and VI-24. The definitions of these procedures may be found in section (iii). The distributions for EXTRAP, described in section (ii), were based on the same model, but a slightly different set of sample cases (i.e. data was not interpolated, because it was not necessary to do so with this program); see Tables VI-8 and VI-9 for the original figures.

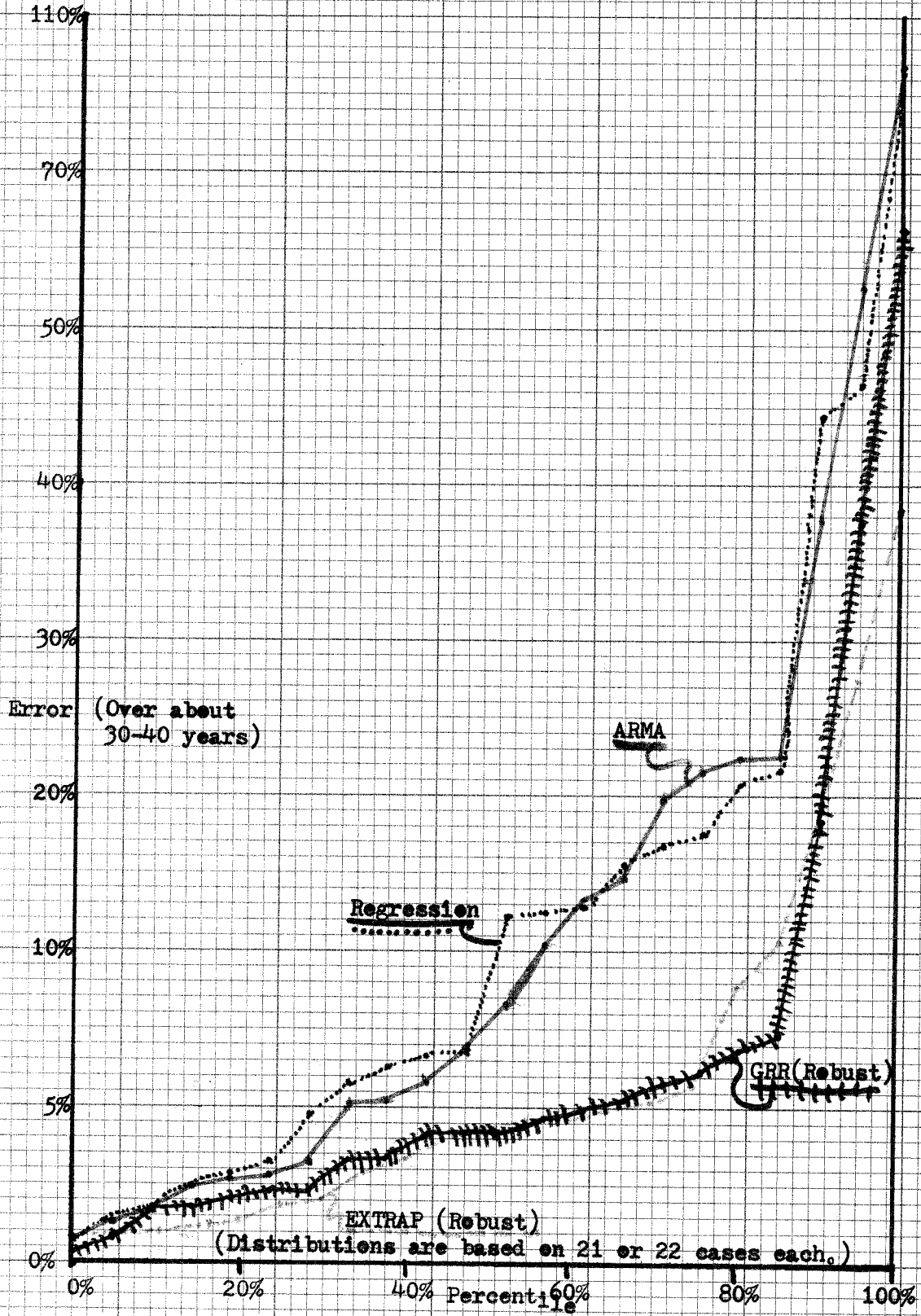

Figure VI-1: RMS Average Errors In Long-Term Predictions of Assimilated Populations, in Percentages. See Legend, P. VI-3.

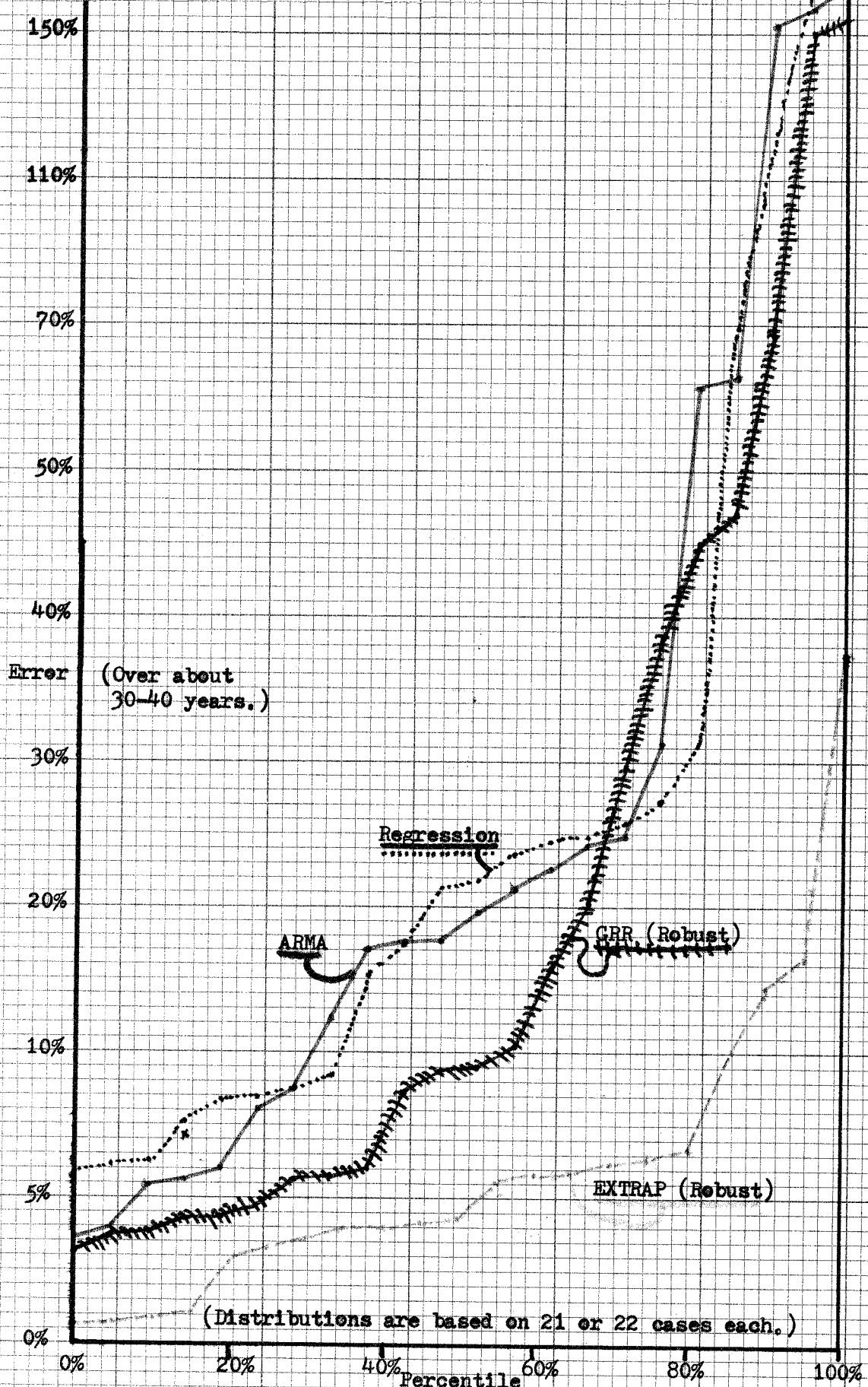

Figure VI-2: RMS Average Errors In Long-Term Predictions of Differentiated Populations, in Percentages. See Legend, Page VI-3.

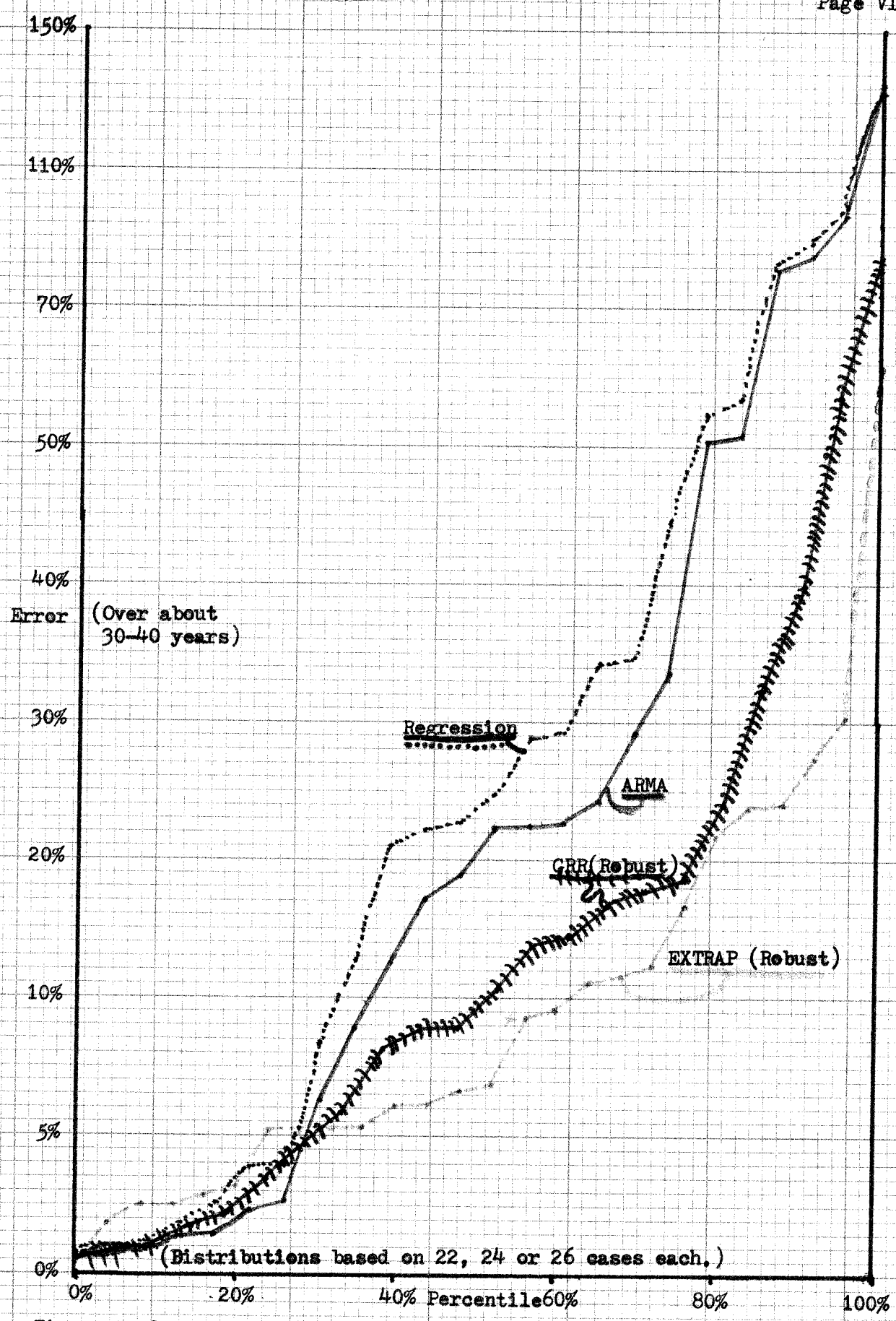

Figure VI-3: RMS Averages Errors in Long-Term Predictions of Mobilized Populations, in Percentages. See Legend, Page VI-3.

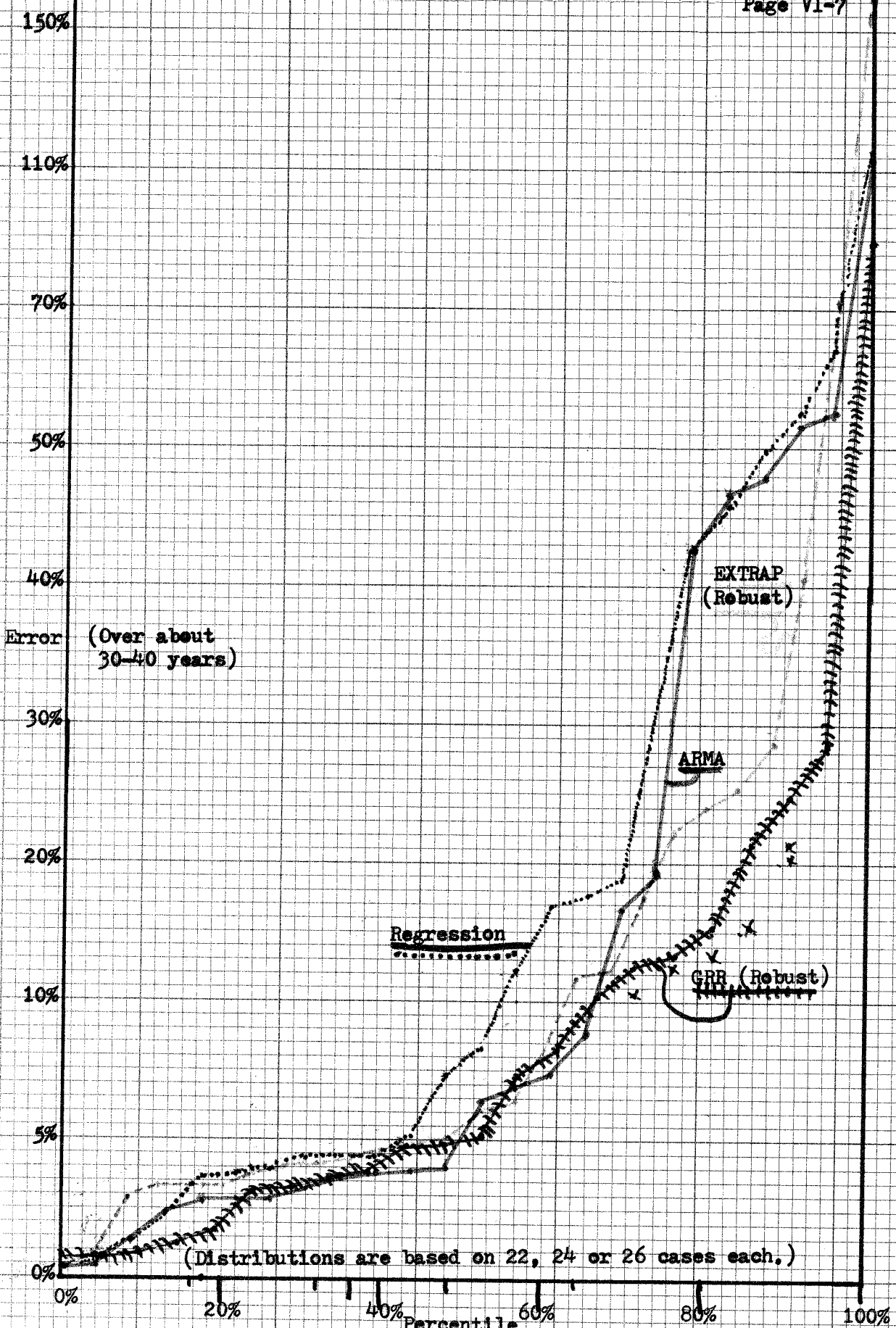

Figure VI-4: RMS Average Errors In Long-Term Predictions of Underlying Populations. See Legend, Page VI-3.

appear to indicate that prediction errors are cut in half, for prediction over about five or six "units" of time, following whatever "units" - years, decades or five-year intervals - were used in the original data collection. (Note that all nations were studied for which sufficient data were available from Deutsch and Kravitz, who, in turn, limited their collection effort only by the requirements that a nation must have a significant problem of national assimilation and that the data must be easily available in the Harvard libraries.) Although these indications have been rather strong, one should still be warned that strict statistical generality and uniformity have not been possible in this case, due to the limited supply of data per nation and due to limitations of existing computer software routines; thus the use of judgement is required to interpret these results. Chapter (IV), by contrast, has been written to provide a test of the various methods which avoids such real-world difficulties. (See Table IV-1, on page IV-34.) In Figures VI-1 through VI-4, we have graphed the distribution of prediction errors for the three methods tested by comparable procedures over the Deutsch-Kravitz data:(i) classical regression; (ii) the ARMA technique of Chapter (III); (iii) the robust

method. In Figure IV-1, which gives the curves for predicting the assimilated population - for which the comparison between methods is exact - one can see that the robust method yields errors distributed uniformly between 0% and 7%, except for three or four (i.e. probability 20%) "fluke" cases. In predicting the percentage of population assimilated - which is not directly affected by random population loss in war or the like - Table VI-9, on page VI-23, shows a uniform error distribution with the robust method between 0% and 2%, with four "flukes" at 2.68%, 3.08%, 3.09% and 6.21% errors. Insofar as much of the data here has been encoded in terms of decades, these prediction errors refer to periods of time on the order of thirty or forty years, for the most part.

In Tables VI-21 and VI-22, we have tabulated the predictions of the robust method for assimilation and mobilization in the countries studied, for the years 1980, 1990 and 2000. It must be emphasized, however, that the model used in generating these predictions - unlike the more complex models discussed below - is not suitable for evaluating the effects of policy in changing what the numbers will equal in those years; furthermore, these predictions require care in their interpretation, as we will discuss in section (iii).

In sections (ii) and (iii), we will elaborate on the details of the statistical methods which led to all these tables and graphs. Also, in keeping with the general philosophy expressed in Chapter (V), we will mention some substantive impressions and hypotheses which emerged after inspecting the predictions generated by the robust approach; a more quantitative study of these hypotheses may be appropriate for future research.

The second strand of our research into nationalism was more substantive in nature. We went back to the original reasoning of Karl Deutsch, in Nationalism and Social Communications(2), and formulated a model of the assimilation process which more fully articulates the vision expressed in that book; this model, by describing the forces which can speed up or slow down national assimilation, can be of direct value to the policy-maker who wishes to do one of these two things. Such a more complete articulation, however, required the addition of intranational "communications terms," whose evaluation in turn required detailed data on assimilation and communications data at a subnational level. Furthermore, the models involved were strongly nonlinear; before we can carry out valid long-term prediction based on these models, our work elsewhere in

this thesis suggests that we will have to wait until the full nonlinear version of the "robust method" is available in a standard computer package such as T.S.P. (Indeed, the full analysis of interaction terms between comparable subunits across time also requires a novel approach to statistical data management as well; however, the Janus subsystem at the MIT-Harvard Cambridge Project may be able to overcome this further difficulty.) Insofar as communications terms are actually necessary, to account for the changing rates of assimilation crucial to Deutsch's verbal discussion, we have concluded that reliable predictions based on that discussion will have to await another round of research. In the meantime, however, we have used the standard regression techniques, and the new ARMA techniques discussed in Chapter (III), in order to evaluate our new communications approach, in a test case - Norway - for which intranational data were very plentiful.

According to the conventional measure of statistical significance, the "ARMA" communications model, based on migration data as an index of communications, outperformed all the other models so well that there is less than one chance in a million billion that this superiority was due to a coincidence.

The conventional measure of statistical significance depends on short-term predictive power, on the quality of predictions over one "unit" of time. (In this case, one "unit" means one year ahead into the future.) In long-term prediction of the percentage of population assimilated, the ARMA communications model made errors which were only about 10% less than the errors with the best competing models. (In this case, "long-term" errors are the average errors over all possible intervals of prediction from one to thirty years into the future; the averages are based on the root-mean-square, "RMS", averaging procedure, which places greater weight on the largest errors and thus on the longest prediction intervals.) However, given the large, diverse data base available, this 10% reduction would appear to be just as significant as the reduction in error in short-term prediction.

In section (v), by looking at the final two studies of the Norway data, and also by comparing them with our earlier, less perfect studies of the same data, we have been able to extend further our discussion of the regression method and the ARMA method as such. In particular, the significance scores of ARMA models do indeed seem to be more sensitive to the quality of the substantive part of the model and to the

quality of the data involved (i.e. flukes) than are those of regression models. On the other hand, the ARMA method shows only a modest improvement over regression in reducing the size of errors in long-term prediction (again, about a 10% reduction in error size), particularly if one's model and data are good to start with; this reinforces our conclusion, from sections (ii) and (iii), that the estimation of model parameters, for use in the long-term prediction of real social data, is better done by way of the "robust" approach.

Finally, in our study of Norway data, we have looked into the possibility of using "gravity models", to reconstruct the networks of interregional communication for years in which no data were available. These models, in standard form, do not allow one to express or explain the changes in communications patterns which underly the effects of modernization, as discussed by Deutsch and interpreted in section (iv). Thus we have generalized the simple gravity model of internal migration, in order to remedy this defect, and have established the validity of this generalization.

(ii) INITIAL STUDIES OF THE DEUTSCH-SOLOW MODEL

Before plunging into the mathematical details of our work on the Deutsch-Kravitz data, let us begin with a general description of the studies we have carried out. These studies fall into two subcategories. In this section, we will discuss our initial studies, in 1971, in which our sole objective was to make use of existing methods and existing models - primarily the Deutsch-Solow model - in order to make concrete predictions of national assimilation and political mobilization. Given that we did not expect to generate or evaluate any new methods as such, and given that the software packages available to us then did not allow long-term forecasting for a linear model structured like the Deutsch-Solow model, we set up special purpose computer programs of our own, tailored to this particular problem alone. Each of these programs included an estimation part, to estimate the constants of the Deutsch-Solow model for each nation individually, and a prediction portion, to make long-range predictions. Three programs were used, based on three methods of estimation, in order: (i) the Hopkins programs(3) ;(ii) standard regression; (iii)the "robust method."

Nation and Base Years	Predicted Values Year	Mobil.	Assim.	Actual Values Year	Mobil.	Assim.	"%Error" Mob.	Ass.
Ceylon	1921	1471	2791	1921	1537	2770	4.29	0.76
1891,1901,'11	1951	3444	4166	1953	4509	5209	23.6	20.0
Thailand	1958	19926	...	1960	18381	...	8.4	...
1925*,'36,'47								
Malaysia	1951	891	3045	1947	835	2428	6.7	25.4
1911,'21,'31	1961	1258	4064	1957	1425	3126	11.7	30.0
Ceylon	1885	602	1302	1881	394	1698	52.8	23.3
1919*,'36*,'53	1902	885	1815	1901	773	2141	14.5	15.2
Scotland	1815	0	...	1821	697	...	100.	...
1881,1814*,'47								
USSR	1970	101331	...	1965	121600	...	16.7	...
1928,'42*,'56	1914	0	...	1914	25800	...	100.	...
Malaysia	1905	188	994	1911	193	1368	2.59	27.3
1931,'44*,'57	1918	269	1316	1921	292	1569	7.88	16.1
Canada	1901	3138	2653	1901	2014	3711	55.8	28.5
1941,'51,'61	1931	4838	5670	1931	5572	7000	13.2	19.0
Philippines	1900	870	1205	1903	1003	3219	13.3	62.6
1936*,'48,'60	1924	2960	3287	1918	3139	3977	5.7	17.4
Quebec	1901	1972	746	1901	645	1212	206.	38.5
1941,'51,'61	1931	2058	1903	1931	1814	2292	13.5	17.0
Argentina	1915	5863	5012	1914	4157	5511	41.0	9.05
1930*,'45*,'60	1900	5575	1857	1892*	1857	2733	200.	32.1
Czechoslovakia	1930	5396	5960	1930	7850	7340	31.3	18.8
1900,'10,'20*	1940	4828	5588	1937	8020	7500	39.8	25.5
India	1911	30210	294756	1911	28482	217197	6.07	35.7
1881,'91,1901	1941	0	0	1941	49792	270187	100.	100.
Argentina	1936	17743	10879	1930*	6914	8625	157.	26.1
1870*,'92*,1914	1958	79683	24650	1960	14758	17440	440.	41.3

Table VI-1: Sample of Results from DELTA
All data measured in thousands of people.
Asterisk represents estimates from ESTIMATES.
Mobil.= mobilized population; Assim. = assimilated;
exact definition of both in Table VI-10.
"%Error" is a crude percentage, from figures listed.

Nation	Data-base	r_M	g	R_M	f	S.E. of f	P of f
Ceylon	1811-1921	.999	.834	1.000	-.26	.241	.3
Cyprus	1881-1931	.999	.996	.999	-.082	.095	.4
Taiwan	1960-1966	.999	.999	.999	.051	.146	.7
USSR	1950-1965	.999	.703	.999	.484	.206	.02
Finland	1800-1960	.996	.988	.996	.002	.030	.9
USA	1790-1960	.995	.998	.995	.168	.112	.15
Japan	1920-1940	.440	.307	.995	2.73	.310	.01
Finland	1958-1967	.993	.976	.994	.581	.532	.3
Quebec	1851-1961	.992	.917	.994	1.13	.745	.15
"	1901-1961	.981	.885	.993	3.03	1.27	.05
Finland	1880-1960	.992	.965	.992	.036	.128	.75
Canada	1851-1961	.989	.983	.991	-.85	.632	.2
USSR	1922-1931	.986	.991	.990	.281	.167	.1
USA	1880-1960	.986	.989	.987	1.03	1.24	.45
Canada	1901-1961	.974	.992	.983	-1.96	1.56	.2
Belgium	1880-1930	.952	.318	.972	-4.54	3.85	.3
India	1881-1941	.969	.902	.970	-.1	.2	.7
Japan	1920-1960	.923	.519	.948	.874	.576	.15

Table VI-2: Regression Statistics for Mobilized and Underlying Populations. "f" is the estimate of f_2, the rate of mobilization, in the regression model, (6.6). The "S.E. of f" is the standard error of f, the conventional measure of the likely size of errors in the estimate of f. "P of f" is the probability that an estimate of f this large, or larger (of either sign) would have happened by coincidence, for an f which is zero, according to conventional theory. "g" is the estimate of g_2, the natural growth factor of the underlying population, minus the rate of mobilization, in equation (6.6); it is also a close approximation to the autocorrelation of the size of the underlying population. R_M is the multiple correlation coefficient between the predictions of (6.6) for mobilization and the actual values; r_M is the autocorrelation of mobilization. Data definitions in Table VI-10; cases listed in order of R_M here.

Nation	Data-Base	r_A	c	R_A	b	S.E. of b	P of b
Taiwan	1956-1965	1.000	.997	1.000	-.031	.065	.6
Quebec	1901-1961	.998	.938	1.000	-1.7	.406	.01
Cyprus	1881-1931	.999	.991	.999	-1.91	1.41	.2
USA	1790-1960	.998	.971	.999	3.489	.927	.01
Israel	1951-1967	.998	.969	.998	-1.29	1.20	.3
USA	1880-1960	.994	.934	.998	9.939	3.11	.01
Finland	1880-1960	.996	.464	.996	-.57	.69	.45
Taiwan	1946-1965	.994	.999	.996	.104	.049	.05
Israel	1951-1960	.994	.999	.994	.795	3.90	.85
Canada	1901-1961	.972	.981	.991	.815	.326	.05
Belgium	1880-1930	.915	.915	.942	.391	.421	.4
India	1881-1941	.863	.976	.922	2.44	1.67	.15
Ceylon	1881-1921	.814	.994	.847	.376	.862	.7

Table VI-3: Regression Statistics for Assimilated and Differentiated Populations. "b" is the estimate of b_2, the assimilation rate, in the regression model, (6.5). The "S.E. of b" is the standard error of b, the usual measure of the likely size of errors in the estimate of b. "P of b" is the probability that an estimate of b this larger or larger (of either sign) would have happened by coincidence if b is actually zero, according to conventional theory. "c" is the estimate of c_2, the natural growth factor of the differentiated population minus the assimilation rate as in equation (6.5). R_A is the multiple correlation coefficient, between the predictions of (6.5) for assimilation and the actual values; r_A is the autocorrelation of assimilation. Data definitions in Table VI-10; cases listed in order of R_A here.

Nation	Data-Base	Base Year	Median Error Model	Median Error Uni.	Median Size	Pct. Error Model	Pct. Error Uni.
Taiwan	1960-1966	1960	4	5	3340	.120	.150
Cyprus	1881-1931	1901	.18	.12	48	.375	.250
Ceylon	1881-1921	1901	19	9	773	2.46	1.16
India	1881-1941	1901	792	697	28500	2.78	2.45
Belgium	1880-1930	1900	171	53	3846	4.45	1.38
Finland	1880-1960	1880	31	40	642	4.83	6.23
Canada	1901-1961	1911	149	524	5572	2.67	9.40
Japan	1920-1960	1930	3739	2680	20022	18.7	13.4
USA	1790-1960	1830	1550	1780	12000	12.9	23.2
Finland	1800-1960	1800	193	123	642	30.1	19.2
USSR	1950-1965	1965	64000	1000	78500	81.5	1.28
Canada	1851-1961	1871	2609	347	2644	98.7	13.1
Japan	1920-1940	1930	10336	5482	20022	51.6	27.4
Quebec	1901-1961	1911	703	211	710	99.0	29.7
USA	1880-1960	1880	9850	4610	12000	82.1	38.4
USSR	1922-1931	1924	19000	huge	78500	24.2	huge
Finland	1958-1967	1958	824	607	642	128.	94.5

Table VI-4: Long-Term Prediction Errors with SERIES In Predicting Mobilization. In each case, i.e. a row in this table, the coefficients of equations (6.6), the "full" Deutsch-Solow model, and of the Univariate ("Uni.") form of this model (i.e. with the "fU" term removed), were estimated by regression over the "data-base". Then, starting from real data in the "base year", predictions were made to all years for which we had data in that case. (See Table VI-10 for data definitions.) The median of the prediction errors, across the years, is shown in thousands in the two columns on the left, first for the full model, then the univariate; from the median mobilization, we have calculated a rough "median percentage error," whose limitations and downwards bias are mentioned in the text.

Nation	Data-Base	Base Year	Median Error	Median Size	Pct. Error
Taiwan	1960-1966	1960	4	8562	.047
Canada	1901-1961	1911	11	4805	.229
Belgium	1880-1930	1900	40	3184	1.26
Finland	1880-1960	1880	35	2529	1.38
Ceylon	1881-1921	1901	39	2793	1.40
Cyprus	1881-1931	1901	3.9	208	1.88
India	1881-1941	1901	5785	274518	2.11
Canada	1851-1961	1871	102	3645	2.80
Finland	1958-1967	1958	81	2529	3.20
USA	1880-1960	1880	1760	33000	5.33
Finland	1800-1960	1800	150	2529	5.93
Quebec	1901-1961	1911	75	1017	7.37
Japan	1920-1940	1930	3906	46358	8.43
USA	1790-1960	1830	4970	33000	15.1
Japan	1920-1960	1930	7220	46358	15.6
USSR	1950-1965	1965	39000	109200	35.7
USSR	1922-1931	1924	45560	109200	41.7

Table VI-5: Long-Term Prediction Errors with SERIES In Predicting the Underlying Population. In each case, i.e. a row in the table, the constants of equations (6.6), the Deutsch-Solow model were estimated by regression over the "data-base." (Note that the "univariate" version is the same as the full Deutsch-Solow model, in predicting the underlying population.) Then, from real data in the "base year", predictions were made to all years for which we had data in that case. (See Table VI-10 for data definitions.) The median of the prediction errors, across the years is tabulated in thousands; from the median size of the underlying population, we have calculated a rough "median percentage error," whose downwards bias is mentioned in the text.

Nation	Data-Base	Base Year	Median Error Model	Median Error Uni.	Median Size	Pct. Error Pct. Error Model	Pct. Error Pct. Uni.
Taiwan	1956-1965	1955	9	9	899	1.00	1.00
Cyprus	1881-1931	1901	2	3	199	1.01	1.51
Israel	1951-1967	1960	21	25	1859	1.13	1.34
Israel	1951-1960	1953	25	37	1859	1.34	1.99
Quebec	1901-1961	1911	41	52	2270	1.81	2.29
Finland	1880-1960	1880	51	61	2754	1.85	2.21
Taiwan	1946-1965	1955	14	22	899	1.56	2.45
Belgium	1880-1930	1900	56	76	2705	2.07	2.81
India	1881-1941	1901	3531	9170	216249	1.63	4.24
Ceylon	1881-1921	1901	56	32	827	6.78	3.87
Canada	1901-1961	1911	582	205	5381	10.8	3.81
USA	1790-1960	1830	7960	5338	38496	20.7	13.9
USA	1880-1960	1880	253207	3660	38496	658.	9.51

Table VI-6: Long-Terms Prediction Errors with SERIES In Predicting Assimilation. In each case, i.e. a row in this table, the coefficients of equations (6.5), the "full" Deutsch-Solow model, and of the Univariate form of this model (i.e. (6.5) with the "bD" term removed), were estimated by regression over the "data-base." Then, starting from real data in the "base year", predictions were made to all years for which we had data in that case.(See Table VI-10 for data definitions.) The median of the errors in prediction, across the years, is shown in the two columns on the left, in thousands, first for the full model, then the univariate; from the median values of assimilation, we have calculated a rough "median percentage error," whose downward bias is discussed in the text.

Nation	Data-Base	Base Year	Median Error	Median Size	Pct. Error
Israel	1951-1960	1953	1	230	.434
Ceylon	1881-1921	1901	29	2739	1.06
Taiwan	1946-1965	1955	87	8175	1.06
Taiwan	1956-1965	1955	101	8175	1.24
Cyprus	1881-1931	1901	1	57	1.75
India	1881-1941	1901	1542	85803	1.80
Belgium	1880-1930	1900	83	4336	1.91
Israel	1951-1967	1960	5	230	2.17
Finland	1880-1960	1880	11	393	2.80
Quebec	1901-1961	1911	19	605	3.14
Canada	1901-1961	1911	403	4996	8.07
USA	1790-1960	1830	920	6491	14.2
USA	1880-1960	1880	5563	6491	85.7

Table VI-7: Long-Term Prediction Errors with SERIES in Predicting the Differentiated Population. In each case, i.e. a row in the table, the constants of the Deutsch-Solow model, (6.5), were estimated by regression over the data-base. Then, from real data in the base year, we made predictions for all years for which we had data in that case. (See Table VI-10 for data definitions.) The median of the prediction errors, across the years, is tabulated in thousands (here, univariate predictions are same as full model predictions); from the median size of the differentiated population, we have calculated a rough "median percentage error", whose downwards bias is mentioned in the text.

Nation	Data-Base	Mobil. Error	Under. Error	%Mobil. Error
Taiwan	1960-1966	.4%	.4%	.17%
India	1881-1941	9.5%	3.4%	.57%
Ceylon	1881-1953	2.8%	3.4%	.57%
USSR	all*	2.5%	4.1%	.80%
Malaysia	1911-1957	5.1%	3.3%	.81%
C.S.S.R.	1900-1937	2.5%	2.9%	.83%
Belgium	1880-1931	5.3%	.9%	1.09%
Israel	1952-1967	1.8%	4.4%	1.18%
Finland**	1880-1960	6.1%	6.1%	1.86%
Canada	1901-1961	6.6%	4.6%	2.01%
Quebec	1851-1961	6.8%	5.1%	2.18%
USSR	1950-1965	3.2%	6.5%	2.23%
Argentina	1869-1960	12.2%	17.8%	2.28%
Quebec	1901-1961	5.1%	4.1%	2.36%
Canada	1851-1961	11.0%	4.9%	2.38%
USA	1790-1960	61.7%	28.8%	2.72%
Cyprus***	1881-1960	21.8%	8.3%	3.73%
Philippines	1903-1960	6.1%	3.9%	3.77%
Finland	all*	9.3%	11.8%	3.79%
USA	1880-1960	11.5%	25.6%	3.98%
Finland	1958-1967	5.2%	340.3%	5.08%
Japan	1920-1960	16.6%	12.2%	6.38%
USSR	1922-1931	30.2%	72.7%	8.03%
Finland	1800-1960	23.8%	24.0%	8.18%
Scotland	1821-1961	24.0%	40.6%	10.05%
Japan	1920-1940	27.3%	22.2%	10.24%

Table VI-8: RMS Average Errors of Predictions of Mobilized and Underlying Populations, by EXTRAP. In each case (except **), we have given the average of the percentage errors as averaged across predictions to <u>every year</u> for which we had data. (See Table VI-10 for data definitions.) The three columns, in order, give:(i) average of the percentage errors in predicting mobilized population; (ii) the average of percentage errors in predicting underlying population; (iii) the average of the <u>absolute</u> errors in predicting the <u>percentage</u> of population which is mobilized. Cases listed in order of the latter.

* - union of all data-bases shown in this table.
** - test years <u>only</u> include data-base proper.
*** - Errors less than 2% uniformly for runs made over early data, used by SERIES.

Nation	Data-Base	Assim. Error	Diff. Error	%Assim. Error
Japan	1948-1965	.9%	4.3%	.03%
Scotland	1891-1963	4.4%	4.0%	.14%
Israel	1951-1966	1.0%	1.1%	.16%
India	1881-1941	4.0%	4.1%	.37%
Israel	1951-1959	1.4%	3.6%	.39%
Canada(B)	1931-1961	4.6%	.9%	.57%
Cyprus	1881-1960	2.0%	6.4%	.61%
Quebec(A)	1901-1961	2.9%	6.6%	.75%
Belgium	1880-1947	4.6%	3.3%	.80%
Finland	1880-1960	1.9%	10.5%	1.04%
Canada(A)	1901-1961	5.2%	2.8%	1.10%
USA	1790-1960	27.3%	16.4%	1.11%
Ceylon	1881-1963	8.8%	6.3%	1.12%
Quebec(B)	1931-1961	.8%	5.8%	1.22%
C.S.S.R.	1900-1937	1.1%	5.8%	1.41%
Philippines	1903-1961	5.7%	4.0%	1.68%
Malaysia	1911-1957	3.3%	5.5%	2.05%
Taiwan	1956-1965	18.1%	.7%	2.68%
Taiwan	1946-1965	38.3%	.7%	3.08%
USA	1880-1960	10.7%	14.5%	3.09%
Argentina	1869-1960	5.0%	37.2%	6.21%

Table VI-9: RMS Average Errors of Predictions of Assimilated and Differentiated Populations, by EXTRAP. In each case, we give the average of the percentage errors, averaged across predictions to <u>every year</u> for which we had data. (See Table VI-10 for data definitions.) The three columns, in order, give: (i) average of the percentage errors in predicting assimilated population; (ii) the average of the percentage errors in predicting the differentiated population; (iii) the average of the <u>absolute</u> errors in predicting the <u>percentage</u> of population which is assimilated. Cases listed in order of the latter.

Argentina:M=Urban.(b); A=Ethnicity (average of series).

Canada:M=Urban.; A=English-Speaking Only (DELTA);
=Ethnicity British Isles (SERIES);
=Not French Ethnicity (EXTRAP A);
=Not French-Speaking Only (EXTRAP B).

Ceylon:M=Literacy; A=Buddhist, except in SERIES, where A=Hindu.(Comparison of univariate models still possible by symmetry.)

Cyprus:M=Urban.; A=Greek Orthodox (SERIES);
=Not Moslem (civilian) (EXTRAP)

CSSR=Czechoslovakia:M=Urban.; A=Ethnicity Czech. (Deutsch estimates;Bohemia,Moravia,Silesia only.)

Belgium:M=Urban.; A=French-Speaking Only

Finland:M=Urban.; A=Finnish-Speaking

India:M=Urban.; A=Hindu. Deutsch population estimates.

Israel:M=Urban.; A=Total Jews

Japan:M=Urban.; A=Not Korean. 5-year data interval.

Malaysia:M=Urban.; A=Malayan Ethnicity

Philippines:M=Literacy; A=Visayan Ethnicity.

Scotland:M=Urban.; A=Speaks No Gaelic

Taiwan:M=Urban.; A=Mainland Chinese

Thailand:M=Literacy

USA:M=Urban.; A=White

USSR:M=Urban.

Table VI-10: Indices of "M"(Mobilization) and of "A"(Assimilation) From the Deutsch-Kravitz Data Used For Runs Reported in Tables VI-1 to VI-9. "Urban." means "Urbanization."

With each of the first two programs, it required no more than a glance at the computer outputs of predicted versus actual values, and of the estimated values of the constants, in order to see that something was not working according to plan. (See Tables IV-1, and VI-4 through VI-7. The latter group of tables seem to imply that assimilation was much easier for SERIES to predict than was mobilization. However, in the actual year-by-year printouts, the difference was less; the use of "median error" appears to exaggerate the difference.) It looked as if the main objective of these early studies - the construction of reasonable predictions - was not going to be possible. A more detailed examination of the regression statistics convinced us that bad estimates of the constants of the Deutsch-Solow model were at fault, not the model itself, for the large size of the errors; the statistics also hinted very strongly that the bad estimates might be due to the random inaccuracies - "measurement noise" - which afflict all normal sources of data in the social sciences.

At this point, we were very lucky to be unable to do what we wanted to do, and to be restricted to a method which our recent work has shown to be much better. We wanted to carry out estimation based on the

Deutsch-Solow model, with terms added to reflect the presence of "white noise" in data collection, in addition to terms reflecting randomness in the process itself; however, we were able to account for "white noise" in data collection only by dropping one substantive term in the Deutsch-Solow equations, and by dropping the terms which allow for randomness in the assimilation and mobilization processes themselves. Thus we unintentionally were using a special case of the "measurement noise only" method, the "robust method" discussed in sections (vii) and (xi) in Chapter (II). This special case is essentially equivalent to an advanced form of curve-fitting and extrapolation. The predictions of this method were quite good, as we have discussed in the Introduction and illustrated in Figures VI-1 through VI-4; the graphs for this method are based on Tables VI-8 and VI-9, on Pages VI-22 and VI-23. Having found a method suitable for our purpose, we then modified our computer program to calculate root-mean-square average percentage errors for the predictions of this model.

Now let us look more closely at the mathematics of these three sets of studies. All three studies were based upon variations of the revised version of the Deutsch-Solow model(4). This model includes two

equations for the process of national assimilation:

$$\frac{dA}{dt} = aA + bD$$

$$\frac{dD}{dt} = cD, \qquad (6.1)$$

where A represents the assimilated population and D represents the differentiated population. The first of these equations states that the rate of growth of the assimilated population, A, with time, is equal to the sum of two different terms. The first of these terms, "aA", refers to the natural growth of the assimilated population through births and deaths. It is assumed that "a" is effectively constant, for our purposes; in other words, it is assumed that births and deaths average out to a fixed percentage of the population itself. The second of these terms refers to increases in the assimilated population, due to unassimilated people being assimilated. It is assumed that "b" is effectively constant; in other words, it is assumed that the number of people assimilated per unit of time averages out to be a fixed percentage of the number of unassimilated people still available. Finally, in the bottom equation, a single term - "cD" - is enough, mathematically, to express the total effect of both

such assumptions on the growth rate of the differentiated population. Also, the "uniqueness" of different countries is acknowledged by acknowledging that "a", "b" and "c" will be different in different countries. The Deutsch-Solow model for political mobilization is virtually identical:

$$\frac{dM}{dt} = eM + fU$$

$$\frac{dU}{dt} = gU, \qquad (6.2)$$

where M is the mobilized population, U the underlying population, "e" the natural growth rate of the mobilized population, "f" the rate of mobilization as a fraction of the underlying population, and "g" a constant analogous to "c", above. Deutsch(5) and Hopkins(6) have discussed in great detail the ability of the model to capture the essence of certain portions of the history of nationalism. In section (iv) of this chapter, we will suggest ways in which larger aspects of this history may be captured by extending this model; in that section, and in Chapter (II), we discuss ways in which we can go beyond the initial simplifying assumption that "a", "b" and "c" are constant.

Mathematically, equations (6.1) imply that there

exist constants a_1, b_1 and c_1 such that:

$$A(t+1) = a_1 A(t) + b_1 D(t)$$

$$D(t+1) = c_1 D(t). \qquad (6.3)$$

This time, instead of talking about the instantaneous rates of growth of population, we are talking about the total growth over one unit of time, from time "t" to time "t+1". The actual unit could be a year, five years, a decade, or anything else we choose. "a_1" represents the natural factor of increase of the assimilated population over one unit of time; thus, an annual population growth of 3% per year would imply $a_1 = 1.03$. "b_1" represents the fraction of the differentiated population which are assimilated per unit of time, adjusted slightly for their own natural increase during the period in which they are assimilated. "c_1" represents the natural growth factor of the differentiated population minus the fraction of the people assimilated. In a similar manner, equations (6.2) imply that:

$$M(t+1) = e_1 M(t) + f_1 U(t)$$

$$U(t+1) = g_1 U(t) \qquad (6.4)$$

In order to make actual predictions of assimilation and mobilization data by use of equations (6.3) and (6.4) above, the major substantive problem is that of estimating the values of the "constants" a_1, b_1, c_1, d_1, e_1 and f_1. (In order to be more precise, instead of calling these things "constants", we will refer to them hereafter by the mathematical term, "parameters"; a "parameter" is assumed to be fixed <u>within</u> a given process - e.g. assimilation in one nation - but may vary from process to process and may also be treated as a kind of unknown variable.) In our initial work in 1971, we used three different methods to estimate these parameters.

First of all, we tried to use the Hopkins method(7). The Hopkins method is based on the assumption that equations (6.3) and (6.4) are <u>exactly</u> true, for the measured values of the variables in every country, for the right values of the parameters. Thus if we know D(t+1) and D(t) for some time t, then we can <u>solve</u> for c_1, as an unknown, in the bottom equation of (6.3). In a similar way, we can solve for all the other parameters , by use of simple algebraic equations, if we know the values of A, D, M and U at three consecutive times. In order to solve for these

parameters, and to carry out predictions on the basis of the resulting estimates, we simply made use of the original Hopkins programs, DELTA and ESTIMATES. Fourteen runs were made, on data from eleven nations, selected from the Deutsch-Kravitz data. The results are shown in Table VI-1. These predictions appear to be extremely poor; a quick examination of the table will make it clear why further effort was not invested in this approach. Indeed, from the point of view of a statistician, as discussed early in Chapter (II), one would not expect to achieve much success with the false assumption that (6.3) or (6.4) are exactly true. Even more emphatically, one would expect that the use of all the data available would give us better estimates of the parameters, than would a series of only three time-points, given that the model may be "true" only in a statistical sense.

Our next step was to estimate the parameters a_1, b_1, c_1, e_1, f_1 and g_1, by use of the classic statistical method, by multiple regression. More precisely, we fit a regression model of the form:

$$A(t+1) = a_2A(t) + b_2D(t) + k_1 + n(t)$$

$$D(t+1) = c_2D(t) + k_2 + m(t), \qquad (6.5)$$

where "n(t)" and "m(t)" are error terms which we try to minimize, and where the parameters "a_2", "b_2" and "c_2" have the same interpretation as "a_1", "b_1" and "c_1". The "constants terms", k_1 and k_2, were added because they tend to be standard in regression studies; otherwise, however, this model is essentially just another way of interpreting the Deutsch-Solow model, (6.1) or (6.3). In like manner, we fit a regression model of mobilization:

$$M(t+1) = e_2M(t) + f_2U(t) + k_3 + n(t) \qquad (6.6)$$

$$U(t+1) = g_2U(t) + k_4 + m(t)$$

In addition, we computed a number of standard regression statistics, to go with these models. These included the "autocorrelations" of A, D, M and U (e.g. the correlation coefficient of A(t) against A(t+1)); they included the <u>probability</u> of the proposition that b_2 and f_2 - the rates of assimilation and mobilization - might be zero, as measured by standard statistical significance tests; they included the correlations ("multiple R") between the actual and predicted values of A(t+1); etc. Statistics of this sort are usually reported as the final results of studies on

quantitative political science, as if they themselves were conclusive. Thirty-one runs were made, on twelve nations, again on the Deutsch-Kravitz data. The results are summarized in Tables VI-2 and VI-3.

Looking at these tables, we can see that the values of R, the correlations between the predictions of our model for A(t+1) or M(t+1) and the actual values of A(t+1) or M(t), tend to be very close to 100%. Thus one would expect these regression models to have unusually great predictive power. Also, b_2 and e_2, the rates of assimilation and mobilization in (6.5) and (6.6), tend to be very large; this hints that all terms of the Deutsch-Solow model are justified and maesurable, quantitatively. However, the autocorrelations, r_M and r_A, also tend to be very large. Given the short length of the data series, this would imply that there is not much information (residual variation) here about those components of A(t+1) and M(t+1) which cannot be predicted from knowledge of A(t) and M(t) alone; indeed, the "standard errors" of b_2 and f_2 turned out to be very large, implying large expected errors in the estimation of these parameters. Even so, in a number of cases, b_2 and f_2 were significantly different from zero, despite the large standard errors, enough to validate the

importance of the cross-terms, $b_2 D(t)$ and $f_2 U(t)$. (Note that "significantly different from zero" means that there was a low <u>probability</u> that they could actually be zero, according to the usual significance measure. When b_2 is estimated to be large, but the estimation error appears to be larger yet, then the true value of b_2 might just as well equal zero.) When b_2 or f_2 is not significantly different from zero, but still apparently large, and when both R_M and r_M or R_A and r_A are near to 100%, one would expect <u>both</u> the regression model <u>and</u> the autocorrelation model (i.e. (6.5) with the bD term removed, or (6.6) with the fU term removed) to have unusually great predictive power.

For our purposes, however, it was not enough simply to look at the regression statistics. The multiple correlation coefficient, R_A, is a reasonable measure of our ability to predict A(t+1) from knowledge of A(t) and D(t), with the help of equations (6.5); however, true long-range prediction entails the ability to predict A(t+T) from A(t) and D(t), <u>without</u> knowledge of intermediate values of A and D, for time-differences, "T", which may be very large. In order to carry out and test such predictions, based on equations (6.5) or (6.6), for all possible values of "T", we wrote a special-purpose program, SERIES, in

FORTRAN. In each test case, for assimilation or mobilization in one country, SERIES computed the standard regression statistics, shown in Table VI-2 or VI-3, and estimated the parameters of model (6.5) or (6.6). The model (6.5), with the error terms removed, corresponds to a unique real differential equation of the form:

$$\frac{dA}{dt} = a_3 A + b_3 D + k_5$$
$$\frac{dD}{dt} = c_3 D + k_6 \qquad (6.7)$$

Note that these equations are essentially the same as the original Deutsch-Solow equations for assimilation, (6.1), with only a couple of constant terms added. SERIES would begin by using regression to estimate a_2, b_2, c_2, k_1 and k_2 in (6.5); then, by solving the corresponding equations, (6.7), it could calculate the values of a_3, b_3, c_3, k_5 and k_6 corresponding to those estimates; finally, by using its solution of (6.7), it could then predict A(t+T) and D(t+T), for <u>any</u> T, even T which are not whole numbers, from the initial data, A(t) and D(t), at any time-period, t. For comparison purposes, SERIES also carried out a parallel set of estimations and predictions, based on (6.5) and (6.7)

with the bD terms removed. This corresponds, in effect, to assuming that assimilation is proportional to the population already assimilated, when we try to predict the assimilated population. The procedures to predict M(t+T) and U(t+T), based on (6.6), were essentially identical to those described above for A(t+T) and D(t+T).

Thirty-one runs were carried out with SERIES, based on data from twelve countries, selected from the Deutsch-Kravitz data. In each country, SERIES actually printed out the prediction and error for every individual year; however, it would be impossible to reproduce all that output here. Thus in Tables VI-4 through VI-7, we have listed the median errors, in numerical and percentage form, for each run. From a formal statistical point of view, Tables VI-15 and VI-18, to be discussed in section (iii), gives a more standard measure of the validity of the regression method as such. However, these earlier results do retain some interest.

Looking at Tables VI-4 through VI-7, we can see that SERIES did at least a plausible job of prediction. The median errors run to ten to fifteen percentage points, for prediction periods on the order of a half-century. The contrast with Table VI-1 is quite

clear. In only one case - the case of mobilization in the USSR - do the Hopkins programs appear to outperform SERIES, in predicting one test-year (1920). However, in that case, the model used by SERIES was fitted to a data-base much further away from the test year, relative to the length of the data-base itself, than was the data-base used with the Hopkins routines; furthermore, the data base used with the Hopkins routines, in this case, included only one time point on the early side of World War II, a war which appears to have had a major effect in perturbing the population of the USSR. A perfect comparison, of course, is impossible, since the Hopkins routines by nature require a more limited data-base than that of SERIES; however, the overall performance of the Hopkins routines, over the cases tested, was clearly inferior to the overall performance of SERIES. With both regression models - the full model, and the model with b_2 or f_2 removed - the predictions made from a smaller data-base held up fairly well over a later test-range, in comparison with predictions made from a longer data-base.

On the other hand, the full model (including b_2 or f_2) performed worse, not better, than the reduced model. If a better estimate of b_2 or f_2 would improve

the predictions of the models (6.5) or (6.6), then one must conclude that the estimates produced by regression are worse than the estimates produced by arbitrarily setting b_2 and f_2 to zero; if we believe, from apriori knowledge and from Tables VI-2 and VI-3, that b_2 and f_2 are substantially different from zero, and that the data do give us some knowledge about this difference, then we must conclude that regression does a poor job of accounting for the existing evidence regarding the values of these parameters. Insofar as a simple model, like (6.5) or (6.6), is too difficult and complex for classical regression to handle, then the development of more complex and more realistic models would indeed require new techniques. Furthermore, the predictions of both the simple and complex models, while reasonable, were not nearly as good as the high values of "R" and "r" seemed to portend. Therefore, in attempting to generate good predictions, we decided to waste no more effort on this fruitless approach.

Finally, in the third of our initial studies, we considered the possibility that the deficiencies of the regression models were due to measurement noise problems, as discussed in Chapter (V). Indeed, in Tables VI-2 and VI-3, one can see that the autocorrelations - r - do not seem to be much higher

for data measured by years than for data measured by decades; this implies that the predictability of the underlying process in not much less over longer time intervals, and that the gap between the observed correlations and a perfect correlation of 100% may be largely due to data measurement error.(8). If the "ARMA" techniques discussed in Chapter (III) had been available at this time, to fit all the parameters in the full model, (6.3) and (6.4), we would have used them; fortunately, however, we had no choice but to use a different method, which has turned out to be superior.

In order to make some allowance for measurement error, we found ourselves forced to eliminate the cross-terms in the original Deutsch-Solow model, equations (6.1), to get:

$$\frac{dA}{dt} = a_4 A$$
$$\frac{dD}{dt} = c_4 D \qquad (6.8)$$

In predicting the differentiated population, this is equivalent to the original model; in predicting the assimilated population, this is equivalent to assuming that the number of people assimilated per unit of time

is proportional to the number already assimilated, and that a_4 incorporates the sum of the effects of natural growth and assimilation. We found it necessary to assume that these equations are exactly true, for the true, underlying values of "D" and "A". By making these strong assumptions, we were able to cope with the possibility that the recorded values of A(t) and D(t), which we will call A'(t) and D'(t), are different from the true values, A(t) and D(t). We can express this possibility by writing:

$$A'(t) = A(t) + m(t)A(t)$$
$$D'(t) = D(t) + n(t)D(t), \qquad (6.9)$$

where m(t) and n(t) are random error terms which we try to minimize. Notice that we decided to minimize the measurement errors as a percentage of the true values, rather than minimize their absolute values; when the population values grow by a large factor, it seems reasonable to expect that the absolute size of the measurement errors will grow along with them. This model, which makes allowance for measurement noise only, is a simple application of the "measurement-noise-only" approach, the "robust

approach" discussed in section (vii) of Chapter (II). In this simple case, for m(t) and n(t) much less than one (e.g. about 10%), the use of this model reduces to the use of sophisticated curve-fitting. More exactly, equation (6.8) and (6.9) imply a close approximation(9) to the simple regression models:

$$\log A'(t) = k_7 + a_4 t + m(t)$$
$$\log D'(t) = k_8 + c_4 t + n(t), \qquad (6.10)$$

where k_7 and k_8, like a_4 and c_4, are parameters to be estimated, defined as:

$$k_7 = \log A(0)$$
$$k_8 = \log D(0) \qquad (6.11)$$

After using simple regression to estimate these parameters, we may go on to use equations (6.10) to predict A and D at other times:

$$A(t) = e^{k_7 + a_4 t}$$
$$D(t) = e^{k_8 + c_4 t} \qquad (6.12)$$

The procedure used in predicting mobilization by this method is exactly analogous. For the third of our

three initial studies, we wrote a FORTRAN program, EXTRAP, to fit such a model to the Deutsch-Kravitz data; forty-seven runs were carried out over seventeen nations, as shown in Tables VI-8 and VI-9. The large number of runs were possible because EXTRAP, unlike SERIES or DELTA, did not require that the data-base used in fitting the model involve measurements at regular intervals only; equations (6.10) do not refer to a standard interval of time-separation, between "t" and "t+1".

The results from using EXTRAP are shown in Tables VI-8 and VI-9, on Pages VI-22 and VI-23; the root-mean-square ("RMS") average percentage errors in predicting the various populations - assimilated, mobilized, underlying and differentiated - have been graphed, and shown in Figures VI-1 through VI-4.

From these graphs, it is clear that EXTRAP performs surprisingly well. In the case of assimilation, the prediction errors were uniformly distributed between 0% and 7% in 80% of the cases; in 20% (four) of the cases, they were much larger. A case-by-case reexamination of the outlying 20% suggested to us that unusual factors - war, or chronic depopulation for economic reasons - were at work on all components of the population, in these cases; for

example, Russia, Japan and Cyprus, when predicted from before World War II to after, were prominent among the outliers. Thus we suspected that the percentage of assimilation given by the model might be substantially more reliable in such cases. Indeed, when we tabulated the error in predicting the percentage assimilated or mobilized, shown in the rightmost columns of Tables VI-8 and VI-9, the performance of the model looks still better. In the case of assimilation, there was a uniform distribution of error between 0 and 2% in 80% of the cases; in the remaining four outliers, the percentages of error were 2.68%, 3.08%, 3.09% and 6.21%. Looking at the choices of "data-base" and "base year" indicated in these tables, we can see that these predictions were made over fairly long intervals of time; in 25 out of 47 cases, the total interval, from the earliest data-base year to the last test year, was at least 60 years. It strikes us as significant, however, that the mobilization process is less predictable than the assimilation process by these methods; one might suspect that mobilization is more easily influenced by the fluctuations of variables not accounted for in any of these simple models, variables such as economic development, and that it is less rigidly governed by inertia.

The performance of EXTRAP was not only good; it was substantially superior to the performances of the regression method and of the ARMA method.(10). In Figures VI-1 through VI-4, the error distribution of EXTRAP is substantially lower than the distributions of all other routines, except for "GRR". GRR is a slightly altered form of EXTRAP, written to allow an exact comparison of the robust method against the other two methods - regression and ARMA - in the more recent phase of research. The curves for the ARMA and regression method, shown in these graphs, are based upon the same reduced form of the Deutsch-Solow model, equations (6.8), that were used with EXTRAP; the details will be mentioned in the next section. The curves for the robust method - EXTRAP and GRR - are lower or equal to the other curves essentially across <u>all</u> of the probability distributions, from the worst 10% to the best 10%. On the whole, they look about one-half the size of the other curves, in true area; they are particularly low in the critical region, from the fortieth to the eightieth worst percentiles, in which prediction errors are large enough to cost heavily to the decision-maker, but normal enough that they can be reduced.

In comparing Tables VI-8 and VI-9 against

Tables VI-3 through VI-7, one can also see that EXTRAP was superior to our old regression procedure, SERIES, which was based upon the full Deutsch-Solow model. The contrast is particularly graphic when one inspects the computer outputs of predicted versus actual values for individual years. Unfortunately, these outputs are far too lengthy to be included here; thus we must content ourselves to note that the comparison between EXTRAP and SERIES was <u>consistent</u> with the pattern which has been established more objectively by the more recent studies of section (iii). Tables VI-8 and VI-9 use a "high" measurement of error, which tends to place greater weight on the largest errors in one's sample; Tables VI-4 through VI-7 use a "low" measurement of error, median error. Thus the superiority of EXTRAP is greater than indicated by a direct comparison of the tables. (SERIES would have been expanded, to include a printout of R.M.S. average error, if its predictions had been competitive enough to justify further work.) Still, Tables VI-4 and VI-6 do allow us to see that the <u>reduced</u> form of the Deutsch-Solow model, equations (6.5) and (6.6) with the "bD" and "fU" terms removed, performs better than the original form of these equations, with regression, when the <u>same</u> measure of error is used; thus the inferiority of the former to

the robust method, as shown in section (iii), implies an even greater inferiority of the latter. Also, a direct comparison of the printouts indicated a similar or worse performance by this early regression procedure, relative to that of the newer regression procedure which we have described in our graphs; it indicated an inferiority to EXTRAP by at least a factor of two. For example, in Table VI-9, EXTRAP looks especially bad in predicting the US differentiated population; however, looking at the printout from EXTRAP, for 1790-1960, a median error of 230 shows up, versus 920 for SERIES. As percentages of the median size of the differentiated population, these numbers correspond to 3.5% and 14.2% respectively. In Table VI-7, SERIES looks especially good in predicting the differentiated population of Finland; in the printouts, however, EXTRAP shows a median error of 8.0, versus 11 for SERIES. (i.e. 2.0% and 2.8% errors, respectively.) Thus Tables VI-8 and VI-9 provide a stiff test of the ability of EXTRAP to predict the future, over long time-intervals. They include several tests of prediction from a model fit to one data-base, to later and earlier sets of data.

From the substantive point of view, it is especially interesting to note what sort of situations

have been hardest for EXTRAP to deal with. Japan, Cyprus and the USSR have the largest recorded errors. An inspection of the actual printouts suggests that they all fit the growth patterns predicted by EXTRAP quite well, except for a break-point in World War II. At World War II, the curves shifted by a constant factor, but, except for the resulting change in scale, they seemed to continue on as before the war. This situation is reminiscent of a linear system, affected by a "delta function impulse" (i.e. a transient shock), in the mathematics of engineering and physics; the effect of the shock is to move the system abruptly from one configuration to another, but the same dynamic equations continue to govern the system after the shock as before it. Prof. George E. Box, in a brief visit to Harvard, mentioned to us that his group is working on a form of "intervention analysis" which would be suitable for the statistical study of such discontinuities.

The errors in Scottish "mobilization" appear related to the much-discussed "rural depopulation" of Britain, an issue comparable to the issue of Appalachia in the US. The errors in Scotland appear to depend on the inclusion of data from a full century and a half, a period encompassing different economic trends. Errors were also large, on occasion, when a short data-base

was used to develop a model for predicting over much larger intervals of time (e.g. in Finland); this observation reinforces our emphasis on using a data-base which is large in actual time, as discussed in Chapter (V). Looking at these two extreme cases - Finland and Scotland - one would be tempted to suggest that the ideal length of the data-base, per case of data, is somewhere in the middle, somewhere on the order of only twice the interval of time over which one is trying to predict. However, as in section (vii) of Chapter (II), we must distinguish between <u>qualitative</u> improvements in one's models, and <u>quantitative</u> estimation of the coefficients of a model which has already been specified and which one knows to be oversimplified. Here we are dealing with the latter problem; with the former problem, we suspect that the longest possible data-base would be desirable.

In two other cases - the USA and Argentina - moderately large errors may be related to changes in both birth rates and death rates in different ethnic groups, during the periods studied; as with Scotland, one might consider these errors symptomatic of too long a data-base, for the simple model under consideration.

(iii) LATER STUDIES OF THE DEUTSCH-SOLOW MODEL

Let us begin our discussion here from a general point of view, as in the previous section, and hold back the mathematical details until after the overall pattern is clear.

After the work discussed above, it was clear to us that we had run up against a general methodological problem, which goes well beyond the requirements of the Deutsch-Solow model itself. Therefore, in order to cope with this problem, we wrote a new computer routine, "ARMA", discussed at length in Chapter (III), for inclusion in a standard computer statistical package for the social scientist (i.e. the Cambridge Project "Time Series Processor", "T.S.P."). According to classical maximum likelihood theory, the basis of our arguments in Chapter (III), this routine should have been the answer to the problem of simple "measurement noise." Included in this routine was a provision to test the long-term predictive power both of a regression model and of the corresponding ARMA model. Generality, however, required us to remove the special-purpose differential equation solving used in SERIES and in EXTRAP, a provision which had allowed us to cope more exactly with the Deutsch-Solow model and

Nation and Data-Base	f_1	LogP $f \neq 0$	LogP ARMA Max.	LogP ARMA Uni.
USSR (1)	.006	2.4	11.4	11.8
USSR (3)	.030	4.2	8.7	9.8
Argentina (2)	.612	6.1	8.2	9.0
C.S.S.R.	.316	.6	10.2	9.0
Malaysia	-.046	6.8	5.7	8.3
USA (2)	.083	.5	7.9	8.3
Cyprus	.315	2.9	12.6	8.1
India	-.142	3.8	7.4	7.0
Philippines	.087	.2	6.7	6.6
USA (1)	.027	1.0	7.0	6.5
Ceylon	-.013	.3	6.2	6.1
Taiwan	-.044	1.7	10.6	6.0
Canada	-.035	.1	4.8	4.8
Israel (2)	.292	7.8	4.0	4.6
USSR (2)	-.047	1.6	4.2	4.5
Finland (1)	-.157	.3	4.5	4.4
Quebec	.005	.0	4.5	3.8
Argentina (1)	.860	1.9	7.3	3.5
Finland (2)	-.008	.4	8.0	3.5
*** LogP (ARMA Uni.) Significant Above This Line ***				
Israel (1)	.493	5.6	6.3	1.8
Finland (3)	-.037	1.2	4.9	.4
Japan	.419	.1	3.2	.4
Belgium	.164	3.1	4.6	.07

Table VI-11: Statistics Concerning Regression for Mobilized and Underlying Populations.
"f_1" is the value of "f_1", the rate of mobilization, in equations (6.4), as estimated by ordinary regression. "LogP" is the standard, classical statistical measure of the relative likelihood of one model in comparison with another; it represents the natural logarithm of the odds in favor of the truth of the model we are interested in, compared with some other model, if we assume that both models had an equal chance of being true apriori.(See section (v), Chapter (V).) Thus if LogP is 4.6 or more, then the odds are better than 100 to 1 that our model is better. In the first of these columns, we compare equations(6.4) against (6.14), to get the probability that f is not zero; in the second, we compare the ARMA "Maximum" model (i.e.(6.18) adapted to mobilization) against the regression version of the model, to see if ARMA is better; in the third, we compare the ARMA version of (6.14)("Univariate") against regression. For LogP of 6.9 or more, odds are 1000 to 1 or better; for LogP of 3, 20 to 1.

Nation and Data-Base	b_6	LogP $b \neq 0$	LogP ARMA Max.	LogP ARMA Uni.
Ceylon	-1.853	6.8	8.4	11.6
Argentina	.977	1.9	9.2	7.8
C.S.S.R.	.137	.4	10.7	7.1
Malaysia	.311	1.5	7.9	6.9
USA (2)	2.698	9.0	9.6	6.3
Finland	.038	0.0	6.2	5.9
Canada	1.945	3.1	8.3	5.7
USA (1)	2.054	3.1	3.4	5.3
Israel (1)	4.215	14.3	7.5	5.3
Scotland (2)	8.343	1.9	4.4	4.8
Scotland (1)	1.471	1.1	4.3	4.7
Quebec	.365	2.4	7.2	4.3
Israel (2)	3.963	13.1	3.9	3.7
India	1.466	1.5	5.9	3.3
*				*
* LogP (ARMA Uni.) Significant Above This Line				*
*				*
Philippines	-.627	.9	4.4	2.6
Taiwan (4)	-.077	1.7	2.2	2.1
Taiwan (3)	.021	2.1	1.9	1.8
Cyprus	-.507	3.2	.5	1.3
Taiwan (2)	-.136	1.4	4.4	.8
Taiwan (1)	-.107	1.1	.8	.8
Belgium	-.220	2.8	5.7	.5

Table VI-12: Statistics Concerning Regression for the Assimilated and Differentiated Populations. "b_6" is the value of "b_6", the rate of assimilation, in equation (6.15), as estimated by ordinary regression. "LogP" is the standard, classical statistical measure of the relative likelihood of one model in comparison with another; it represents the natural logarithm of the odds in favor of the truth of the model we are interested in, compared with some other model, if we assume that both models had an equal chance of being true apriori.(See section (v), Chapter (V).) Thus if LogP is 4.6 or more, then the odds are better than 100 to 1 that our model is better. In the first of these columns, we compare equation(6.17) against (6.17), to get the probability that b is not zero; in the second, we compare the ARMA "Maximum" model against the regression version(6.18), to see if ARMA is better; in the third, we compare the ARMA version of (6.16) and (6.17)("Univariate") against regression. For LogP 6.9 or more, odds are 1000 to 1 or better; for LogP of 3, 20 to 1.

Nation and Data-Base	f	LogP f≠0	Rho Mobil.	Rho Under.
Israel (1)	.499	10.1	1.63	.182
Belgium	.164	7.7	-1.13	.03
Cyprus	.232	7.4	-1.19	-1.424
Israel (2)	.325	7.2	.746	.204
Finland (3)	-.037	6.7	-1.517	.317
Taiwan	-.044	6.3	-1.641	1.710
Malaysia	-.046	6.2	-1.518	-1.440
Argentina (1)	.852	5.8	-1.329	1.819
Argentina (2)	.613	5.3	-1.72	1.780
Finland (2)	-.012	4.9	-1.009	.733
India	-.142	4.2	-1.687	-1.447
Japan	.089	3.1	.325	.064
USSR (3)	.030	3.1	.565	1.296
*				*
* LogP Significant Above This Line				*
*				*
USSR (1)	.007	2.0	.650	.712
C.S.S.R.	.316	1.8	-2.204	-2.257
USA (1)	.030	1.7	.629	1.529
USSR (2)	-.047	1.3	.530	-1.085
Ceylon	-.012	.6	-1.096	-1.230
Quebec	.000	.6	-1.25	-.201
Philippines	.083	.4	-1.517	-1.643
Finland (1)	-.158	.4	-1.27	-.162
USA (2)	.08	.2	.64	1.196
Canada	-.036	.1	.360	1.229

Table VI-13: ARMA Models for Mobilization Processes. "f_1" is the rate of mobilization in equations (6.4), as estimated by the ARMA method. "LogP", as in Table VI-11, is the classical measure of the probability that f is actually nonzero, despite the uncertainty of estimation; in this table, we use the ARMA models to calculate LogP. LogP here is the natural logarithm of the odds in favor of the proposition that f is not zero. When LogP is 4.6 or more, the odds are 100 to 1 or better that f is nonzero; when LogP is 3 or more, the odds are 20 to 1. "Rho" is a coefficient, discussed in the text, which tends to be nonzero when data collection or other measurement errors are large (rho=1 is very large); when rho is zero, the ARMA model reduces to a regression model.

Nation and Data-Base	b_6	LogP $b \neq 0$	Rho Assim.	Rho Diff.
Israel (1)	4.215	19.7	-.006	-1.15
Israel (2)	3.959	13.3	.583	.441
USA (2)	2.673	12.2	-1.09	1.155
Belgium	-.219	8.1	-1.35	-.323
Canada	1.946	5.7	1.926	1.205
Quebec	.348	5.4	-1.09	1.383
C.S.S.R.	.139	4.1	-1.310	-1.399
India	1.453	4.1	-1.150	1.416
Ceylon	-1.855	3.6	-1.148	1.620
Malaysia	.309	3.5	-2.142	-1.326
Argentina	.977	3.4	3.14	1.761
*				*
* LogP Significant Above This Line				*
*				*
Philippines	-.632	2.7	-1.02	-1.292
Cyprus	-.507	2.3	.043	.336
Taiwan (3)	.021	2.2	.017	.373
Taiwan (4)	-.076	1.7	-.079	.402
Scotland (2)	8.351	1.3	-1.634	-.324
USA (1)	2.042	1.2	-.141	-1.15
Taiwan (1)	-.107	1.1	.799	.339
Scotland (1)	1.597	.7	1.686	.025
Finland	.055	.3	1.407	.057
Taiwan (2)	-.137	0.0	5.0	.360

Table VI-14: ARMA Models of the Assimilation Process. "b_6" is the rate of assimilation in equation (6.15), as estimated by the ARMA method. "LogP", as in Table VI-12, is the classical measure of the probability that b is actually nonzero, despite the uncertainty of estimation; in this table, we use the ARMA models to calculate LogP. LogP here is the natural logarithm of the odds in favor of the proposition that b is nonzero. When LogP is 4.6 or more, the odds are 100 to 1 or better that b is nonzero; when LogP is 3 or more, the odds are 20 to 1. "Rho" is a coefficient, discussed in the text, which tends to be nonzero when data collection or other measurement errors are large (rho=1 is very large); when rho is zero, the ARMA model reduces to a regression model.

Nation and Data-Base	Mobil. (6.18) Max.	Mobil. (6.15) Model	Mobil. (6.17) Uni.	Under. (6.18) Max.	Under. (6.16) Uni.
Taiwan	.26	.23	.73	.19	.84
USSR (3)	1.24	1.28	.85	1.32	1.32
USSR (1B)	.84	.83	.95	.82	4.44
USSR (2)	8.40	7.54	1.88	1.08	.44
Finland (3)	.99	.95	2.58	.70	2.37
Ceylon	5.08	5.27	3.86	3.42	3.73
C.S.S.R.	2.52	2.65	3.91	3.89	3.65
USSR (1A)	1.58	1.58	8.39	.65	4.46
Quebec	8.60	8.22	12.88	8.52	7.42
India	17.45	241.48	21.02	8.09	5.10
Malaysia	8.29	11.66	22.20	5.80	4.43
Japan	33.09	34.64	22.86	29.74	18.98
Israel (1)	7.18	10.28	24.90	24.55	8.31
Belgium	4.29	4.31	28.94	.85	3.99
Canada	93.73	99.59	29.11	10.07	19.43
Philippines	14.60	14.67	34.11	5.69	12.18
Israel (2)	31.56	5.60	34.82	52.89	49.91
Finland (1)	10.02	8.57	44.95	508.77	16.98
Cyprus	17.49	18.00	54.67	6.66	17.62
Argentina (1)	11.61	11.45	57.01	7.23	54.93
Finland (2)	59.99	51.32	83.91	100.64	45.94
USA (1)	63.83	67.97	89.00	388.89	64.05
Argentina (2)	14.09	13.02	100.25	16.86	42.58
USA (2)	47.04	45.85	133.14	91.21	116.66

Table VI-15: RMS Averages of Percentage Errors With Long-Term Predictions of Mobilized and Underlying Populations, Based on Regression. In each case, the four models used were fitted to the "data-base" defined in Table VI-23, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are <u>averages</u>, in each case, across all such test years. The five columns give errors with four different models; these models are the equivalent (i.e. have the same structure, with mobilization switched for assimilation, etc.) of the equations whose numbers are listed in the column headings. "Mobil." means "Mobilized"; "Under." means "Underlying". "RMS" means "Root-Mean-Square" (i.e. Averages taken as the square root of the arithmetic average of the squares.)

Nation and Data-Base	Mobil. (6.18) Max.	Mobil. (6.15) Model	Mobil. (6.17) Uni.	Under. (6.18) Max.	Under. (6.16) Uni.
Taiwan	.43	.22	.72	.10	.49
USSR (3)	1.19	1.18	.86	1.31	1.40
USSR (1B)	.85	.85	1.00	.85	3.64
USSR (2)	8.52	6.82	1.38	1.11	.47
Ceylon	5.22	2.27	1.48	3.43	2.81
C.S.S.R.	2.77	2.63	2.38	3.36	3.77
Finland (3)	.97	.79	2.55	.66	2.41
Quebec	6.72	5.59	6.41	5.22	6.92
USSR (1A)	1.78	1.47	8.92	.48	3.90
Philippines	14.96	11.53	12.60	4.81	2.83
Cyprus	17.53	16.35	17.01	6.66	6.47
Japan	33.37	38.07	18.91	33.74	19.23
India	16.22	55.83	22.22	6.60	3.21
Canada	57.82	93.74	22.42	12.24	8.82
Malaysia	5.70	1.29	22.70	6.50	2.69
Israel (1)	9.34	8.57	24.04	24.09	7.33
Belgium	4.32	3.44	29.05	.87	3.98
Israel (2)	29.24	6.05	33.65	54.85	47.94
Argentina (1)	11.69	8.84	51.11	11.64	46.74
Finland (1)	9.65	11.43	53.63	241.79	16.62
USA (1)	63.76	61.76	81.05	481.04	53.27
Finland (2)	14.29	16.59	84.78	7.04	55.10
Argentina (2)	12.86	7.65	96.40	8.18	42.75
USA (2)	53.79	46.19	135.70	110.62	113.57

Table VI-16: RMS Averages of Percentage Errors With Long-Term Predictions of Mobilized and Underlying Populations, Based on the ARMA Method.
In each case, the four models used were fitted to the "data-base" defined in Table VI-23, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are <u>averages</u>, in each case, across all such test years. The five columns give errors from four different models; these models are equivalent (i.e. have the same structure, with mobilization switched for assimilation, etc.) of the equations whose numbers are listed in the column headings. "Mobil." means "Mobilized"; "Under." means "Underlying". "RMS" means that "Root-Mean Square" averaging was used.

Nation and Data-Base	Mobil. ext1	Mobil. ext2	Under. ext1	Under. ext2
Taiwan	.59	.78	.99	1.49
USSR (3)	.99	1.76	.75	1.49
Finland (3)	1.17	2.42	1.33	2.69
Ceylon	1.82	2.08	3.11	4.01
USSR (2)	2.34	3.42	.77	.96
C.S.S.R.	3.50	4.14	3.83	4.26
Malaysia	4.69	6.61	3.31	4.74
Quebec	5.83	11.47	5.06	9.83
Israel (2)	8.06	29.19	8.46	50.83
India	8.72	10.31	3.64	3.68
Belgium	8.84	15.63	1.83	4.52
Canada	10.42	24.16	4.78	13.40
Finland (2)	13.99	34.76	10.65	21.36
Philippines	14.35	29.90	4.94	9.40
Argentina (1)	16.91	23.10	15.26	24.63
Cyprus	17.95	18.77	7.35	7.39
Japan	18.63	18.96	12.89	17.54
Argentina (2)	24.16	42.12	13.21	21.22
USA (2)	32.25	46.03	24.96	52.04
Israel (1)	39.94	33.84	21.54	12.57
Finland (1)	57.22	53.28	28.20	24.48
USA (1)	85.23	89.70	89.65	84.45

Table VI-17: RMS Averages of Percentage Errors With Long-Term Predictions of Mobilized and Underlying Populations, Based on the Robust Method (GRR). In each case, equations (6.2) with the "fU" term removed were fitted to the "data-base" defined in Table VI-23, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are <u>averages</u>, in each case, across all such test years. "Mobil." means "Mobilized"; "Under.", "Underlying". "ext1" is the variety of robust method used by EXTRAP, described in section (ii); "ext2" is another variety mentioned in section (iii) and tabulated only for the sake of formal completeness. "RMS" means that "Root-Mean-Square" averaging was used.

Nation and Data-Base	Assim. (6.18) Max.	Assim. (6.15) Model	Assim. (6.17) Uni.	Diff. (6.18) Max.	Diff. (6.16) Uni.
Taiwan (3B)	.36	2.98	.73	2.89	7.62
Taiwan (2)	2.96	3.80	1.36	13.35	27.13
C.S.S.R.	1.67	1.35	1.79	6.87	5.92
Taiwan (1)	3.81	5.37	2.48	17.66	25.81
Taiwan (4)	.95	.90	2.85	23.90	161.45
Taiwan (3A)	1.18	25.29	3.23	25.20	166.78
Finland	4.60	4.49	4.82	14.17	17.41
India	6.12	3.91	5.82	7.25	8.80
Scotland (2)	2.97	2.93	6.29	6.89	21.34
Cyprus	1.43	2.14	6.59	4.51	21.94
Scotland (1)	3.34	3.24	6.72	3.26	6.27
Canada	12.59	5.97	12.28	13.84	24.86
Malaysia	3.72	3.81	12.50	3.72	9.18
Belgium	6.47	9.80	12.98	3.58	6.08
Quebec	2.68	2.39	15.64	12.07	31.2
Argentina	41.78	13.64	16.91	145.39	122.23
Ceylon	11.15	192.41	17.49	7.31	15.07
Philippines	27.55	36.40	20.91	22.40	8.47
Israel (1)	1.95	1.95	21.69	8.06	8.37
USA (1)	44.11	5.11	44.40	71.54	69.92
Israel (2)	30.00	3.48	46.56	33.47	23.70
USA (2)	39.04	7.99	97.61	45.61	24.41

Table VI-18: RMS Averages of Percentage Errors With Long-Term Predictions of Assimilated and Differentiated Populations, Based on Regression. In each case, the four models used were fitted to the "data-base" defined in Table VI-24, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are <u>averages</u>, in each case, across all such test years. The five columns give errors from four different models used in predictions; these models correspond to the equations whose numbers are listed in the column headings. "Assim." = Assimilated; "Diff." = Differentiated. "RMS" means that "Root-Mean-Square" averaging was used.

Nation and Data-Base	Assim. (6.18) Max.	Assim. (6.15) Model	Assim. (6.17) Uni.	Diff. (6.18) Max.	Diff. (6.16) Uni.
Taiwan (3B)	.37	2.99	.76	2.97	8.04
Taiwan (2)	3.72	4.81	1.31	13.68	22.64
C.S.S.R.	1.41	1.24	1.78	6.92	5.59
Taiwan (1)	3.52	5.48	2.50	17.61	21.08
Finland	2.91	2.86	2.66	17.65	17.51
Taiwan (4)	.95	.90	2.84	23.93	157.16
Taiwan (3A)	1.18	25.84	3.15	25.24	167.19
Scotland (1)	3.31	4.64	5.08	3.50	6.03
Scotland (2)	2.69	4.09	5.11	15.49	17.05
India	3.47	2.82	5.77	2.72	12.48
Cyprus	1.34	2.11	6.96	4.69	24.37
Malaysia	3.28	3.56	8.24	3.66	3.55
Canada	14.20	5.63	10.32	14.72	17.54
Belgium	4.65	1.99	13.22	2.97	5.48
Quebec	2.31	1.79	14.87	11.97	25.87
Ceylon	10.37	2.78	19.95	8.01	19.61
Philippines	28.01	14.39	21.60	21.81	3.98
Israel (1)	1.76	1.76	22.43	8.30	8.79
Argentina	41.89	13.83	22.67	145.30	151.99
Israel (2)	30.12	3.55	37.86	33.77	31.25
USA (1)	44.03	5.72	55.80	71.45	61.96
USA (2)	40.80	11.05	96.40	46.28	62.30

Table VI-19: RMS Averages of Percentage Errors With Long-Term Predictions of Assimilated and Differentiated Populations, Based on ARMA Methods. In each case, the four models used were fitted to the "data-base" defined in Table VI-24, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are <u>averages</u>, in each case, across all such test years. The five columns give errors from four different models used in prediction; these models correspond to the equations whose numbers are listed in the column headings. "Assim." = Assimilated; "Diff." = Differentiated. "RMS" means that "Root-Mean-Square" averaging was used.

Nation and Data-Base	Assim. ext1	Assim. ext2	Diff. ext1	Diff. ext2
Taiwan (3B)	.37	.66	29.71	53.74
Taiwan (3A)	.75	3.54	47.03	134.42
Taiwan (4)	1.65	3.71	45.04	130.92
C.S.S.R.	1.75	1.79	5.73	7.54
Cyprus	2.03	3.01	6.01	9.85
Taiwan (1)	2.25	1.94	154.54	130.61
Finland	2.27	6.03	9.48	15.89
Scotland (2)	3.30	8.67	10.33	15.48
Scotland (1)	3.42	9.01	3.12	7.74
Malaysia	4.05	6.20	4.71	8.13
Taiwan (2)	4.12	3.69	149.89	125.87
India	4.19	4.25	3.78	4.80
Quebec	4.67	12.15	20.12	31.83
Argentina	4.95	4.96	38.74	66.85
Belgium	5.23	6.67	3.70	3.72
Philippines	5.74	11.44	4.26	4.47
Canada	6.21	14.80	8.61	15.26
Israel (2)	6.92	41.69	4.34	5.82
Ceylon	7.43	7.29	5.63	7.73
USA (2)	18.38	34.22	15.91	35.95
Israel (1)	37.56	29.76	9.41	9.17
USA (1)	62.82	59.97	68.33	69.10

Table VI-20: RMS Averages of Percentage Errors With Long-Term Predictions of Assimilated and Differentiated Populations, Based on the Robust Method (GRR). In each case, equations (6.8) were fitted to the "data-base" defined in Table VI-24, and predictions were made from the "base year" to all later years for which we had data; the errors listed here are averages, in each case, across all such test years. "Assim."=Assimilated; "Diff."=Differentiated. "ext1" is the variety of robust method used by EXTRAP, described in section (ii); "ext2" is another variety mentioned in section (iii) and tabulated only for the sake of formal completeness. "RMS" means that "Root-Mean-Square" averaging was used.

Nation and Data-Base	Mobilization 1980	1990	2000	Underlying Pop. 1980	1990	2000
USA (2)	275.	407.	602.	80.6	95.1	112.
Israel (2)	4.98	9.08	16.6	.572	.638	.713
Finland (2)	3.02	3.80	4.78	2.71	2.94	3.18
Cyprus	.145	.171	.202	.615	.703	.803
CSSR*	11.4	12.3	13.3	2.24	2.11	1.99
Malaysia	3.89	6.01	9.30	7.28	8.70	10.4
Japan	136.	207.	313.	31.1	29.7	28.3
Ceylon	11.2	15.6	21.8	4.15	4.39	4.63
India(+Pak.)*	75.4	84.0	93.4	419.	442.	467.
Taiwan	7.24	11.3	17.8	13.4	17.7	23.2
USSR (3)	255.	419.	687.	141.	170.	204.
Argentina (1)	30.6	44.1	63.6	7.28	8.58	10.1
Argentina (2)	31.5	49.1	76.5	9.86	11.7	13.8
Philippines	30.4	46.4	70.7	18.7	21.6	24.9
Canada	23.5	32.4	44.8	6.49	7.05	7.66
Quebec	6.90	9.31	12.6	1.47	1.54	1.61
Belgium	7.80	8.88	10.1	3.10	3.11	3.12

Table VI-21: Predictions of Future Mobilized and Underlying Populations, by the Robust Method(GRR). "ext2" used; see discussion in text. All figures in millions; definitions in Table VI-23.

* - Base year for predictions was before 1950.

Note that 1974 populations for India, Pakistan and Bangladesh total more than 650 million.

Nation and Data-Base	Assimilation 1980	1990	2000	Differentiated 1980	1990	2000
USA (2)	254.	321.	405.	35.5	42.8	51.6
Israel (2)	4.43	7.12	11.5	.699	1.01	1.53
Taiwan (3B)	14.4	18.5	23.9	45.2	286.	1815
Taiwan (4)	17.5	24.0	32.9	18.5	90.1	439.
Canada	21.7	26.6	32.6	4.53	5.19	5.96
Quebec	7.20	9.16	11.7	.701	.697	.694
Ceylon	9.25	10.9	12.8	4.54	5.22	6.01
Finland	5.02	5.55	6.13	.347	.352	.356
Malaysia	4.67	5.56	6.62	5.64	7.28	9.40
CSSR*	10.1	10.8	11.6	3.08	3.01	2.94
India(+Pak.)*	329.	346.	364.	165.	179.	195.
Cyprus	.638	.742	.863	.128	.142	.157
Scotland (1)	5.43	5.61	5.80	.055	.046	.039
Scotland(2)**	5.45	5.61	5.76	402p	239p	142p
Argentina	30.1	39.5	52.0	4.22	5.41	6.94
Philippines	9.87	25.2	32.0	24.1	30.0	37.1
Belgium	3.44	3.61	3.80	7.43	8.09	8.82

Table VI-22: Predictions of Future Assimilated and Differentiated Populations, by the "ext2" version of the Robust Method (GRR). See discussion in text. All figures in millions, except in Scotland. Definitions of "Assimilated" in Table VI-24.

* - Base year for predictions was before 1950.
** - "p" used to mean "people"; too few for millions.

Nation and Data-Base	Mobilization Definition	Years Model Is Fitted To	Gap In Years	Base Year
USA (1)	urbanization	1790-1870	10	1870
(2)	"	1790-1960	10	1790
Israel (1)	"	1948-1957**	1	1957
(2)	"	1948-1967**	1	1948
Finland (1)	"	1800-1880	10	1880
(2)	"	1800-1960	10	1800
(3)	"	1958-1967	1	1958
Cyprus	"	1881-1961*	10	1901
C.S.S.R.	"	1900-1940*	10	1900
Malaysia	"	1911-1961*	10	1911
Japan	"	1920-1960	5	1920
Ceylon	literacy	1881-1951	10	1881
India(+Pak.)	urbanization	1881-1941	10	1881
Taiwan	"	1960-1966	1	1960
USSR (1A)	"	all below	1	1924
(1B)	"	all below	1	1953
(2)	"	1922-1931	1	1924
(3)	"	1950-1965	1	1953
Argentina (1)	(Table VI-10)	1869-1960*	22.75	1869
(2)	literacy (b)	1869-1960*	22.75	1869
Philippines	literacy	1903-1961*	14.5	1903
Canada	urbanization	1851-1961	10	1851
Quebec	"	1851-1961	10	1851
Belgium	"	1860-1960*	10	1860

Table VI-23: Definition of Mobilization Variables and Spans of Years Used For Runs Described in Tables VI-11 through VI-22. All long-term predictions were made <u>from</u> data in the "base year", up to the end of the continuous string of observations of which it is a part, in the Deutsch-Kravitz data. "Gap In Years" is the interval between observations.

* - interpolated data; N not artificially enlarged.
** - heavily interpolated. (8 years actual data spaced out into 20-year string.)

Nation and Data-Base	Assimilation Definition	Years Model Is Fitted To	Gap In Years	Base Year
USA (1)	White	1790-1870	10	1870
(2)	"	1790-1960	10	1790
Israel (1)	Jewish	1948-1957	1	1957
(2)	"	1948-1967	1	1948
Taiwan (1)	Taiwanese	1946-1955	1	1955
(2)	Not Mainlanders	1946-1955	1	1955
(3A)	Taiwanese	1946-1965	1	1946
(3B)	"	1946-1965	1	1955
(4)	Not Mainlanders	1946-1965	1	1946
Canada	Not French-Only	1901-1961*	10	1901
Quebec	Not English-Only	1901-1961*	10	1901
Ceylon	Buddhist	1881-1961*	10	1901
Finland	Speak Finnish	1880-1960	10	1880
Malaysia	(no choice)	1911-1961*	10	1911
C.S.S.R.	Ethnicity	1900-1940*	10	1900
India(+Pak.)	Hindu	1881-1941	10	1881
Cyprus	Not Moslem	1881-1961*	10	1881
Scotland (1)	Speak No Gaelic	1891-1961*	10	1891
(2)	Speak English	1891-1961*	10	1891
Argentina	Ethnicity	1869-1960*	22.75	1869
Philippines	Visayan	1903-1961*	14.5	1903
Belgium	Speak French	1850-1950*	10	1850

Table VI-24: Definition of Assimilation Variables and Spans of Years Used For Runs Described in Tables VI-11 through VI-22. All long-term predictions were made <u>from</u> data in the "base year", up to the end of the continuous string of observations of which it is a part, in the Deutsch-Kravitz data. "Gap in Years" is the interval between observations.

* - interpolated data; length of data sample not artificially enlarged.

the Deutsch-Kravitz data.

When we applied the ARMA routine to the Deutsch-Kravitz data, using several different versions of the Deutsch-Solow model, we found consistently that:(i) according to the usual measure of statistical significance, the "ARMA" models were indeed superior to the corresponding regression models, with "p" - the probability that this was a mere coincidence - less than .01 in most cases (see Tables VI-11 and VI-12); (ii) in terms of long-term predictive power, the ARMA models did not do very much better than the regression models; they led consistently to a reduction in the size of prediction errors, but only by about 10% of the error sizes at best. (The slight differences in error distributions between the two methods are visible in Figures VI-1 through VI-4.) The second of these two results was also corroborated by our results in Norway, to be discussed in section (v).

There are two immediate corollaries to these results. First, that the usual significance measure is not a good index of long-term predictive power, at least not for models which correspond to the same choices of variables. Second, that the "robust method", which performed much better than regression in our initial research, is superior to the ARMA method as

well, in terms of long-term prediction. (See Figures VI-1 through VI-4, on Pages VI-3 through VI-7.)

In order to document the second corollary more concretely, we have also made use of GRR (GRowth Rate) - a revised form of EXTRAP, for robust estimation in the univariate linear case - to establish an exact correspondence between the three methods, for the assimilation and differentiation data; more precisely, in Figures VI-1 and VI-2, we have graphed the error distributions for all three methods, as methods of making predictions based on the same substantive model, equations (6.8), and as tested over the same sample cases of data drawn from the Deutsch-Kravitz data. In these graphs, the superiority of the robust method is clear, for both the GRR and EXTRAP versions. From the simulation studies of Chapter (IV), one might suspect that this superiority is due in part to the overlaps between the data-bases over which our models are fit and the years over which they are tested. However, we have included a few examples of a time-series split in half, with the model fitted to the first half and the predictions made to the second half. (See Tables VI-23 and VI-24. Israel, Taiwan, USA and Finland are the prime examples, because they are the cases where adequate data was available for such a splitting.)

These examples do not seem markedly different from the other cases studied; unfortunately, these examples are too few to allow a definitive conclusion. We will see, however, that the ARMA model for equations (6.8) has more free parameters to estimate than do either the regression or the robust methods. When the supply of data is very limited, an overlap between the data samples used for fitting and testing would tend to overstate the relative performance of the model with more parameters. (It is standard practice, for example, to try to correct for "degrees of freedom" in one's model, when the data are quite limited; this subject, however, is a Pandora's Box, which we will not open here.) In short, the strong superiority of the robust method over the ARMA method, in these studies, is probably not due to any bias in the details of our procedure.

Before we go more deeply into the mathematical details of these studies, the political scientist might be curious about the projections of the future by the models we have looked at, for some of the countries where they have worked well in the past. In Tables VI-21 and VI-22, we have listed the predictions of the robust method (GRR) for the future, in the countries we have studied. The uses of these numbers

are for the reader to decide for himself; our authorship of the numbers in no way implies that our opinions about their use are any better than anyone else's. However, these opinions are probably worth recording here, at least to provide the reader with a straw man to debate with.

When one first glances at these tables, a few wild numbers immediately grab the eye. How, for example, could the Republic of China (Taiwan) be expected to have 1,800,000,000 inhabitants in the year 2000? (See Table VI-22.) Then, if one has a little serendipity, one will note that only a few million of these are to be "Taiwanese", that about 450 million are to be mainland Chinese, and that the rest would appear to be neither. Visions spring to mind of a collapse of the People's Republic, of a return to the mainland by Chiang's son, and of the expansion throughout the weak nations of Southern Asia by this new, fascistic nationalistic regime. It would be amusing to study the pros and cons of the possibility of such a scenario. However, the simple models we have used do not "know" about (do not account for) the complex factors which might make such a scenario possible; while it is possible for a model to "know" about such factors <u>implicitly</u>, we suspect that these models are too simple

even for that, in the example at hand. A simpler interpretation of these wild numbers is that the predictions of this model, in the future, will be very much like the predictions in the past; they will contain a handful of wild outliers, and a larger number of surprisingly accurate predictions of the percentage of assimilation. It is fortunate that we can spot some of the outliers so easily in advance, simply by using our general knowledge that some of the predictions are absurd.

In the other cases, we would tend to follow the procedures suggested in Chapter (V). We would place greater faith in predictions based on a model estimated over a long data-base - as in Finland - as opposed to predictions based on a lot of data restricted to a shorter period of time (e.g. Taiwan or Israel); indeed, the absurdity of the latter predictions offers some tangible evidence for our point of view. We would try to ask in each case: "What does the model really 'know' about? At what dynamic level could one observe the effects of the forces which will be important in the future? Given those factors which I know about, and given the subset of those which cannot be subsumed under something which the model accounts for, how would I adjust these predictions?" Questions of this sort

lead one to a different approach to applied political science, as we have discussed in Chapter (V).

In some sense, the predictions of these simple models are based on the continuation of the trends which have existed for a long time in the past. It is often comforting for a decision-maker to assume that his administration will somehow be free, at little cost, from the momentum of such trends; indeed, when one is harassed, as most decision-makers are, it is easy to "miss the forest for the trees", and to overestimate the implications of short-term reverse fluctuations. The predictions of this model, in some sense, show the decision-maker what the forest looks like. It is still up to him to use his judgement, to decide whether his administration is truly, objectively likely to perform much differently from those which have dealt with the same problems in the past. Even in concrete terms, these predictions imply no dramatic shift in the percentages of assimilation in the countries studied; a strong upsurge of the Visayans is predicted in the Philippines, and Chinese with recent roots in the mainland are projected to become a majority of the population of Taiwan, but these are the only exceptions.

A few of our readers might also be interested in

the predictions of the regression and ARMA models, in some of the countries where they have worked well in the past. These projections are too voluminous to be duplicated here(11), but their fine details are probably unreliable in any case.

In Canada and Quebec, as a whole, the model predicts little change in the relative balance of French and English speakers, to the year 2000. (This would appear to contradict the separatist claim that the French language would die away without special political measures to bolster it. On the other hand, it implies that French Canadians will remain a political force to be reckoned with. In Canada and in Quebec, our models did better in predicting the longer periods of time, rather than the shorter; the "10% errors" listed in the Tables are mostly from transient deviations from the trend predicted by the models.) A large increase in urbanization is predicted for Quebec, which, in practice, might shift the assimilation trend more to the advantage of French speakers. In Ceylon, a large increase in literacy is projected. In Scotland, a further large decrease in the knowledge of Gaelic is predicted. (If knowledge of Gaelic were a good indicator of political behavior, this would imply that recent signs of a revival of Scottish nationalism are

misleading and transient; however, the connection between linguistic nationalism and political nationalism may not be simple in this case, any more than in the case of Ireland.) In Japan, a huge urban population - 200-odd million - was forecast; however, from the limitations cited in section (ii), it should be emphasized that political and economic factors may cross the threshold of being able to upset this prediction. (In Japan, as in Canada, the errors reported in Tables VI-16 and VI-19 were essentially transient.) In Cyprus and Taiwan, the ARMA models predict little change in the balances between the different factions.

Now let us look more closely at the mathematical details. Our primary interest was in the original Deutsch-Solow model - equations (6.1) and (6.2) - and in the reduced form of this model, with the "bD" and "fU" terms eliminated. (e.g. equations (6.8).) Instead of working with the Deutsch-Solow model directly, in terms of differential equations, we worked with the finite-difference equations which the model implies, equations (6.3) and (6.4); to refresh the reader's memory, let us recall what equations (6.3) looked like:

$$A(t+1) = a_1 A(t) + b_1 D(t)$$

$$D(t+1) = c_1 D(t),$$

where a_1 is the natural factor of growth of the assimilated population, where b_1 is the rate of assimilation per unassimilated person per unit of time, and where c_1 is the natural factor of growth of the differentiated population minus b_1. In like manner, the reduced form of the Deutsch-Solow model, with the "bD" and "fU" terms removed, leads to a reduced form of the finite-difference equations:

$$\begin{aligned} A(t+1) &= a_5 A(t) \\ D(t+1) &= c_5 D(t), \end{aligned} \qquad (6.13)$$

and:

$$\begin{aligned} M(t+1) &= e_5 M(t) \\ U(t+1) &= g_5 U(t). \end{aligned} \qquad (6.14)$$

In predicting the differentiated population, (6.13) is equivalent to (6.3); in predicting the assimilated population, (6.13) is equivalent to assuming that the number of people assimilated per unit of time is proportional to the number already assimilated. It should be emphasized that this reduced form of the

Deutsch-Solow model was studied for the sake of its mathematical simplicity, not for the sake of any hope that it would be superior to the original model on substantive grounds. (In section (v), we will discuss regression models based on what appears to be an intermediate assumption, that the number assimilated can be explained partly as a constant percentage of the overall population; however, the ARMA models to be discussed in that section achieved greater empirical success than these regression models, even though they lacked such a constant term.)

In section (ii), we already described how we applied the "robust method" to the reduced form of the Deutsch-Solow model. Our new routine, GRR, estimates the parameters of that model in exactly the same way, except that it works only on a continuous series of data spaced at regular intervals. In order to use classical regression on equations (6.3) and (6.13), we added an "error term", n(t), to each, and attempted to fit the regression equations:

$$A(t+1) = a_6 A(t) + b_6 D(t) + n(t) \qquad (6.15)$$

$$D(t+1) = c_6 D(t) + n(t) \qquad (6.16)$$

$$A(t+1) = a_7 A(t) + n(t) \qquad (6.17)$$

The terms "n(t)" represent the various random disturbances which we invoke to explain the actual errors we experience in predicting A(t+1) and D(t+1) from the known values of A(t) and D(t), by use of our models. The equations for mobilization were exactly parallel in their structure. For each of the forty-five cases of assimilation and mobilization studied, taken from seventeen nations, each of these three equations was estimated separately by use of the T.S.P. command ARMA. (More precisely, the three equations, in order, were analyzed by issuing the commands: "arma assimilated on differentiatedend", "arma differentiatedend", and "arma assimilated$end". Note that variables which are named after the keyword "on" are treated as "exogenous," as variables to be used in making predictions but not themselves to be predicted by use of the equations at hand.) In addition, since it was impossible to simulate more than one set of equations at the same time, we estimated the set of equations:

$$A(t+1) = a_{\beta}A(t) + b_{\beta}D(t) + n(t)$$

$$D(t+1) = c_{\beta}D(t) + d_{\beta}A(t) + m(t), \qquad (6.18)$$

where "m(t)" is also a random disturbance; this <u>set</u> of equations represents the combination of (6.15) and (6.16), with an extra term added solely for the purpose of creating a "complete set" of equations that fits the "vector ARMA" framework discussed in Chapter (III). (This set of equations was estimated by the command "arma assimilated differentiated$end", or, in the case of mobilization, by "arma mobilized underlyingend".)

In Chapter (III),we have emphasized that there is a correspondence from any regression model, to an ARMA model which says the same thing but which also allows for the possibility of measurement error; for example, (6.15) is equivalent to:

$$A(t+1) = a_9 A(t) + b_9 D(t) + n(t) + Pn(t-1), \qquad (6.19)$$

where "P" may be called a "rho coefficient". Notice that the "rho coefficient" does <u>not</u> multiply a substantive variable in this problem; rather, it multiplies the <u>previous</u> value of the <u>same</u> disturbance. (Elsewhere, of course, we have used the same letter, "n", to refer to <u>different</u> random processes.) From an intuitive point of view, it is simply a measure of the presence of "measurement noise," of collection errors in the available numerical data vis-a-vis the original

underlying concepts, as we have discussed in Chapter (III). When rho is small, this implies that the regrssion model (i.e. the same model without the rho term) is fairly close to the "truth"; more precisely, it implies that tractable aspects of measurement noise will have little effect on one's estimation. When rho is large, this implies that measurement noise is substantial. The four computer commands mentioned above were sufficient, not only to estimate the four sets of regression equations above, but also to estimate automatically the equivalent ARMA equation in every case. For each model estimated, LogP significance scores were printed out; the differences between these LogP scores, for two models being compared with each other, gave us the LogP scores reported in Tables VI-11 through VI-14. The "rhos" reported in these tables for the "Maximum" model (i.e. (6.18) or the equivalent for mobilization) were actually the diagonal terms of the rho matrix, "P", of Chapter (III).

As part of the ARMA command, an automatic simulation facility was also available. After the regression and ARMA models were estimated over a given set of years (a "data-base"), simulations could be made from any given year ("base year") into the future.

Actual data for "endogenous variables", for variables which the model can predict, are taken only from the base year; predictions are made, further and further into the future, by being compounded on other predictions. After the predictions are done, they are checked against real data. Thus the predictions are true long-term predictions. (The percentage errors are calculated as a percentage of the average of the absolute values of the predicted and actual values, and are averaged together by root-mean-square averaging.) However, in equation (6.15), the differentiated population is not internal to the equation; thus the predictions made in that case are not true long-term predictions, for our purposes.

In using GRR, to make and evaluate predictions similar to those made by ARMA, we used two different techniques. One of these techniques, exactly parallel to EXTRAP (and thus to "ext1" discussed in Chapter (IV)), has been used in Figures VI-1 through VI-4; the other, the same as "ext2" in Chapter (IV), starts from the real data in the base year, and uses its estimates of a, c, e and g in equations (6.13) and (6.14) to compound predictions of the future. Error percentages were calculated by the same formula as with ARMA.

In brief, with all of these techniques - ARMA and GRR - data were used, spaced at regular intervals, for fitting the model; then, from a base year, predictions were made to later periods of time, also by regular intervals, with the same formula used to measure percentage error. Given that the original data were not available at regular time intervals, an interpolation routine, "INTS", in TSP was used to create equivalent data at regular intervals, by geometric interpolation; except in the case of Israel, however, the data periods interpolated to were quite close to the original data periods. (In Israel, data were collected annually, but missing data occurred rather randomly, and interpolation was to an annual series.)

The results from these runs are shown in Tables VI-15 through VI-20. We have already discussed the broader implications of these results. A more detailed inspection of the statistics in these tables tends to reinforce those implications, particularly the implications of weakness on the part of the classic maximum likelihood methods (regression or ARMA). Note, for example, that the usual measure of statistical significance gives greater emphasis to the superiority of the ARMA models over the regression models than it does to the values of the cross-coefficients - "b" and

"f" - which represent the rates of assimilation and mobilization. This would seem almost to imply that it is more probable that assimilation and mobilization are not happening, than that the regression model is as good as the ARMA model, if we accept the classical measure at face value.

In terms of minimizing long-term prediction errors, however, the complex, expanded form of the Deutsch-Solow model, equations (6.18), which includes the cross-coefficients, did better than the univariate models, (6.16) and (6.17); this applies to <u>both</u> the ARMA and regression forms of these models, in whatever combination, implying that the cross-terms really were more important in terms of actual prediction errors than the difference between ARMA and regression. Admittedly, however, our prediction tests may have been biased in favor of models with more parameters in them. Still, in acknowledging this bias, one must go on to observe that the univariate robust method - based on far fewer parameters than the ARMA variation of (6.18) - still did better in long-term prediction than either form of (6.18); given that our tests were biased in favor of the latter, the superiority of the robust method is clear.

Also, if we look at the values of the "rho

coefficients", "P", in Tables VI-13 and VI-14, we again see that the standard statistical analyses here come out strongly in favor of the ARMA method. As we mentioned above, the estimated value of rho is a good measure of how different the ARMA model is from the corresponding regression model; the regression model corresponds to the special case where the "rho" coefficients are all set to zero. The rho coefficients do indeed seem to be very different from zero; this would seem to indicate that the processes here strongly require the additional terms provided by the ARMA model. This phenomenon would hint that the mediocre predictive power of the ARMA models may be due to a lack of the quantity of data needed, in each case, to estimate the ARMA coefficients precisely enough. However, the Norway results of section (v) will show that more data per case are not enough to overcome the problem. Also, if we look at the regression and ARMA estimates, both, of the constants "b" and "f" in our models (Tables VI-11 through VI-14), we find many values which look unrealistically high, especially when we stick to the intuitive interpretations of them as "assimilation rates" (e.g. USA, with 267% of all blacks turning white per decade); we know that such assimilation rates are unrealistic, largely because we

know that they would lead to absurd predictions if extended over a few time periods. If we suspect that the true values of these rates would be far smaller, then, according to the usual significance measure, we must admit that the measured values are "significantly different" from zero to about the same extent that they are "significantly different" from their true values. In other words, the error is quite significant; it is not likely to be a coincidence, due to a small quantity of data. Rather, we would say that the error is due to a conventional criterion for likelihood estimation which emphasizes, in practice, only short-term predictive power. If we admit that huge, bad estimates of the cross-coefficients lead to unrealistic predictions when extended over enough time-intervals, then we imply that a different approach to estimation, based on the direct maximization of long-term predictive power, would give us smaller and more realistic estimates. Insofar as the estimates of these coefficients are artificially inflated by the maximum likelihood approach, it is quite possible that the same process affects the "rho" coefficients.

Also, the variability of the signs of the "rho" coefficients tends to imply, from the mathematics at the beginning of Chapter (III), that many of these

large values for the "rho" coefficients are due to something else besides pure measurement noise. In quantitative political science, one often reads statements such as,"Measurement noise with this data would tend to invalidate the regression coefficients; however, such noise would tend to understate the strength of the real connections; therefore, the effects demonstrated here are, if anything, more valid than regression would indicate." If there is a strong possibility of effects which move rho coefficients in the opposite direction from what measurement noise would indicate, then regression may just as easily be overstating the strength of major coefficients. Thus we find, empirically, yet another weakness in the conventional approach to evaluating models.

Finally, from a technical point of view, one may note that almost all of the results in these tables were achieved after ten "major iterations" of the algorithm of Chapter (III). In most cases, convergence proceeded rather steadily, starting out with large movements of coefficient estimates, but proceeding to smaller movements systematically and quickly; convergence was good, most often, after five iterations. In a few cases, however, the total gain in log likelihood relative to regression looked

suspiciously low, when we reviewed the computer output. Ten of the assimilation runs were carried out over again, through many iterations. In all cases but two, it was verified that the routine had indeed converged within ten iterations; indeed, the convergence was generally better than expected (subsequent progress in log likelihood on the order of 0.01), probably because these were cases where the original regression models required little improvement.

On the other hand, there were two exceptions:(i) the application of equations (6.4) (upper equation only) to data on urbanization in Cyprus; (ii) the application of equations (6.18) to data from 1790 to 1870 on white and nonwhite populations in the US. In both cases, the computer printout from the first ten iterations gave a very clear picture of "imbalance", a convergence problem described in section (iv) of Chapter (III). (Very crudely, this problem results from the danger that an estimation system based on first derivatives will be too responsive to some parameters, in comparison with others, and will therefore oscillate so much in response to the former that it makes little headway in dealing with the latter.) In these cases, the general multiplicative factor, used to determine the size of adjustments in

each minor iteration, was decreased, then increased, then decreased, then increased, by very large factors, in a suspiciously regular wave-like pattern. In the other cases of small initial movement, by contrast, the multiplicative factor changed very little after the first few iterations, and changed almost entirely in the downwards direction when it did change. (The adjusted data from Cyprus show up in the Tables. However, the adjusted data for this run in the US do not.) The convergence algorithm we have used has avoided "imbalance" in almost all cases; however, the cases which remain do point up the value of further improvements in convergence procedures, as part of the effort to operationalize the algorithms of Chapter (II). Also, they are worth noting for those who would wish to actually use the command "ARMA" in TSP-CSP.

(iv) NATIONALISM, CONFORMITY AND COMMUNICATIONS TERMS: AN EXTENSION OF THE DEUTSCH MODEL

The original goal of this research was to follow up on the suggestions of Karl Deutsch, in Nationalism and Social Communications, to begin the development of a predictive, quantitative theory of nationalism. These suggestions included a specific mathematical model - the Deutsch-Solow model - which provided the major focus of the work above. They included the suggestion that the use of the dominant national language be used as an index of national assimilation. They also included a number of verbal propositions, presented as "suggestions for future research;" these propositions represent an effort to draw together known verbal relations, bit by bit, into a more coherent dynamic theory, capable of making predictions if only the data were available. By the reasoning of Chapter (V), we believe that this makes them of great substantive interest in their own right. Given that we hoped to exhaust the possibilities of the simple Deutsch-Solow model at an early stage of this research, we have gone back to these earlier propositions, in order to draw them together into a mathematical expression both more

complete and more capable of significant generalization to other problems in political science.

In Chapter 6 of Nationalism and Social Communications, Karl Deutsch presents his main argument that the birth of nationalism may depend on the relative rates of national assimilation and political mobilization, particularly on the latter, more volatile variable.(12). He points out that a moderate-to-slow rate of mobilization will tend to keep the unassimilated groups in the minority, in the cities and in the schools; therefore, those of them who do move to the cities may be assimilated more quickly. A rapid, sudden mobilization, on the other hand, may make the unassimilated groups close to half of the population; they may therefore become more self-conscious as a group, and far less likely to feel the need to assimilate themselves to the old status quo. Conflict may result. The Solow model, which assumes that the rate of assimilation (per unassimilated person) is constant, cannot account for this kind of variation. Thus the Solow model does not articulate Deutsch's critical insight into the origins of nationalism.

In order to express the idea that the rate of assimilation in any area (urban or rural) depends on how much the differentiated are outnumbered by the

assimilated, we may form a model like this:

$$A(t+1) = A(t) + k(A(t) - D(t))D(t), \quad \text{for } A(t) > D(t), \tag{6.20}$$

where A(t) is the <u>percentage</u> of population assimilated at time t, where D(t) is the percentage of population differentiated (i.e. unassimilated) at time t, and k is a constant. In essence, this model states that the <u>percentage</u> of the differentiated who are assimilated in any year will not be <u>constant</u>, as in the Solow model, but will be proportional to the numerical percentage dominance, (A(t)-D(t)), of the assimilated population over the differentiated population. In the first half of Chapter (II), we have discussed a number of refinements which could be made to this simple model. Stanley Lieberson's study of bilingualism in Canada has shown that the effects of local "percentage dominance" are of overwhelming importance(13) in predicting rates of bilingualism and linguistic assimilation; he has shown that fairly smooth curves result when one plots local percentage dominance against language change, implicitly <u>holding constant</u> the overall national linguistic and cultural environment. Even in its simple form, however, equation (6.20) does express the idea of percentage dominance as a determinant of the rate of assimilation.

In order to improve further on equation (6.20), let us consider two other qualitative factors which also need to be accounted for. First of all, let us look at demographic factors. Lieberson, in Language and Ethnic Relations in Canada, has emphasized(14) two competing forces which can affect the fate of a minority language:(i) the economic and percentage dominance of the majority language (English), which encourages people to assimilate away from the minority language (French); (ii) the "revenge of the cradle," the high birthrate of the rural, provincial people who speak the minority language. In this study, we have tried to avoid dealing with the demographic factors directly. By applying (6.20), not to the nation as a whole, but to the urban or rural part of one provincial area at a time, we can expect less difference between the birthrates of the two language groups. If the model is expressed in terms of percentages of people speaking different languages, then the overall birthrates need not be estimated. In order to go on to predict the nationwide percentages of language use, we would have to predict population, first, in each region, and then convert our predictions of percentages of language use in each region to predictions of numbers of language speakers in each region. In brief, this model treats

population growth in each region as an exogenous variable. The primary reason for doing so is simply that this variable has been studied in enormous detail elsewhere, and that it would take enormous work for us merely to duplicate a portion of those studies here; we have looked at the variable, briefly, but we have tried to keep it a different issue, for our limited purposes here.

Second of all, the model, as written, defines "percentage dominance" merely as "A(t)-D(t)", the percentage by which the majority language dominates the minority language. In reality, percentage dominance consists of two different variables - percentage dominance within each locality or region, and percentage dominance nationwide. The first variable encourages regionalism. For example, it encourages people in Quebec to speak French only, while encouraging people in the rest of Canada to speak English only; it makes the regions more and more distinct from each other, and it reinforces the conflict between them. On the other hand, the second variable encourages people in all regions to conform to a national norm.

Local vs. Regional Language-Dominance as a Dynamic Factor.

How could we predict which of these two processes will become dominant, the regional or the national? One way to deal with this question is by trying to formulate an abstract theory of language-dominance pressure. We can try to formulate a theory which does not rely on (arbitrary) political abstractions, like the boundary lines between administrative regions. In the spirit of Nationalism and Social Communications, we can focus instead on the nationwide network of communication flows. For any given individual inside such a network, the language pressure he experiences depends simply on the balance between the two languages as a percentage of his communications, past and future.(15).(These communications should be weighed, in principle, by their psychological salience. Also, the "natural" level of communication between two people or two regions may sometimes be a more accurate measure of language pressure than the actual level, if the latter differs from the former due to a mutual inability to communicate; a desire to communicate, unfulfilled, can sometimes provide an incentive to learn the other person's language.)

Translating this to the level of regional

variables, we may write the model as:

$$P_i(t) = \sum_j C_{ji} \, (A_j(t) - D_j(t)) \qquad (6.21)$$

$$A_i(t+1) = A_i(t) + D_i(t) * f(P_i(t)),$$

where "$P_i(t)$" represents the percentage dominance of the majority language as experienced in region number i at time t, where "C_{ji}" represents the flow of communications between region i and region j (actually, to region number i, from region j, as a percentage of the total communications to region i), where the summation in the upper formula is to be taken over all regions j if possible, and where we have borrowed the asterisk from computer terminology as a sign of multiplication. In the lower formula, we have written "$f(P_i(t))$", instead of just "$P_i(t)$", to reflect our observation above that, in equation (6.20), we could have replaced the expression "A(t)-D(t)" by a more complicated <u>function</u> of the percentage dominance.

Let us look briefly at the implications of this model, for a "typical" nation passing through the stages of political development. Let us suppose that the nation starts out as a "traditional" country, mostly rural, heavily dependent on sedentary agriculture. In such a country, one would expect that

the vast peasant majority would have few communications, if any, outside their own region; the economics of such a region would provide little incentive and little opportunity for the average man to communicate with other regions. On the other hand, the urban and literate subsections of such a country would still have many communications outside their own area, particularly with the merchants and literati of other cities. Our model would therefore predict that regional language pressure would be overwhelmingly dominant over national language pressure, for the illiterate majority. Therefore, the spoken language will sustain a fragmentation into a host of regional dialects(16). On the other hand, the written language of the elite will experience heavy long-distance language pressure, at a national or even international level; it will tend to coalesce into a uniform national or even continental language. All of this assumes a fairly stable and well-divided class structure.

However, as economic growth begins, and rural people become mobilized, a conflict will develop between their original dialects and the national language of the cities they move to. The written language will come into a head-on collision with the spoken system of languages. As the communications

network grows more and more integrated, and more extensive for the average man within each nation, national pressures will grow more important for the average man; thus there is likely to be a growth in national assimilation and uniformity, at the level of the spoken language. On the other hand, if any of the regional dialects "capture" a cohesive center of mobilized population, such as a city, before they are assimilated to the national language (i.e. extremely rapid mobilization occurs), then this city may exert its <u>own</u> language pressure on the surrounding rural population and towns. The widespread, modern communications links acting on this city will <u>strengthen</u> the hold of its dialect, and perhaps lead to a political separatism which then outlaws extensive communications between this region and other regions. If there were one dominant city, such as a London or a Paris, in a large area, then this city, once "captured", may set a new national linguistic norm. If there were a number of competing cities, however, one might expect a greater persistence of local dialects, converging perhaps by a process of mutual adjustment of the language norms themselves, if the norms were close to each other, but not by assimilation as such, until one of the cities does succeed in dominating the

network of communications; thus a greater level of fragmentation would be predicted. Note that the average distance of communications in the new modern network may actually be less than that of the old elite network; thus the new written language may indeed be more restricted, geographically, than the language of the old elite. All of these predictions seem broadly consistent with the phenomena discussed in Chapter 6 of Nationalism and Social Communications.

Finally, one may note that the use of communications terms, as in equations (6.21), can be generalized to other aspects of the problem of nationalism, and even to social psychology on a wider scale. It would be inappropriate, in this context, to discuss all of these future possibilities. However, from a practical point of view, these models are only a small beginning in the quantitative theory of nationalism; much of their value lies in the possibilities that they point to for future research. In order to realize this value, let us sketch out some of these possibilities explicitly.

Some Operational Dimensions of Nationalism: Narcissism, Stereotyping and Aggression.

The Deutsch-Solow model and the models discussed

above involve the origin of nationalism, the origin of systems of national identification. However, it is also interesting to ask how nationalism, once born, can grow to become a motive force behind militarism, chauvanism, and the like. Indeed, one may regard "nationalism" as consisting of seven clusters of variables, only one of which concerns identification directly: (i) affiliation with a "nationality"; concretely, this would entail a clustering at the national level of language norms, cultural symbols, etc.; (ii) the sharing of "tacit norms"(17) that make cooperation possible in situations of mixed conflict and cooperation; when these norms tend to be close or identical among people of the same nationality, but very different for those of different nationality, then "community"(18), by Deutsch's definition, exists precisely at the national level; (iii) the overestimation of the power level of one's own nation ("narcissism", in the language of the psychiatrist); (iv) the underestimation of the power levels of other nations (stereotyping); (v) the intensity of positive emotional commitment to one's own nation (utility attributed to the "success" of one's nation; this may be the resultant of both "rational" and "irrational" (narcissistic or neurotic) attachments); (vi) the intensity of emotional

commitment, positive or negative, to other nations (utility attributed to their "success"); (vii) the glorification of militaristic, nationalistic behavior for its own sake.

"Nationalism," by this definition, would appear to be a crucial cause of international violence. The likelihood of international violence would appear to depend on the ability to compromise in any given perceived game - which we would associate with cluster (ii), above(19) - and in the perception by the participants that compromise is desirable. Insofar as war involves a massive destruction of resources, at least when modern nation-states are involved, one would normally expect it to be far away from what an economist would call "Pareto optimal;" one would expect that a compromise would exist, far superior for <u>both</u> sides than the actual outcome of the war. If the participants overestimate the gains they would achieve by war, however, they may not be able to appreciate beforehand that any particular compromise would be more desirable. Also, if they attach a positive value to hurting their adversary, then they may feel that their own material losses in war would be balanced out by the losses of their adversary. These misperceptions, which may lead to war, are associated with clusters (iii)

through (vi), above. In practice, the breakdown of bargaining between nations will often depend on some "spark", like the assassination of an Austrian archduke or the sinking of the Maine; however, long before the spark appears, there may be a prolonged period of cold war, in which the ability to compromise gradually decreases and peace becomes ever more precarious. Students of conflict who dismiss the sticky, "irrational" factor of nationalism, and focus solely on objective conflicts and capacities, may be helpful in encouraging more objective, less nationalistic and more peaceful policies by the major powers today; however, by neglecting a primary cause of past conflicts, they may reduce the applicability of their historical studies.

One may note, furthermore, that the concept of "nationalism" above is important, not only to the simple variable of war-vs-peace, but also to the possibility of bargains on a higher level, to maximize joint production in ways which would have been impossible without cooperation. Trade agreements are only one part of this picture. In the limit, if nations were fully adept at such negotiation, they could achieve the same joint efficiency and productivity that a unified world government could, if they do not place

a strong negative intrinsic value on each other's basic welfare. On a lower level, such a gradual improvement in coordination has been crucial, in the past, in the fusing of subnationalities into larger nations(20).

Toward More General Models.

In order to predict the variables which make up nationalism, one may try to extend equation (6.20) to deal with continuous psychological variables. Even in dealing with language, we found it necessary at times to talk about changes in the language norms themselves, as continuous variables, rather than simple adherence to one norm or another.(21). Given a continuous variable, X, which represents some arbitrary cultural norm, such as the pronunciation of a certain vowel, we may try to express the idea that an individual will change his own norms in response to the norms of those he communicates with. As a first approximation, we get an equation analogous to (6.20):

$$\frac{dX_i}{dt} = k \sum_j C_{ji} (X_j - X_i) \qquad (6.22),$$

where "X_i" is the value of the variable X (a norm) for person number i, and where C_{ji} represents the strength

of communication between person number i and person number j. In a sense, the term on the right is a "reinforcement" term impinging on person number i.

In practice, however, when we deal with questions like nationalism and fundamental personal values, it is unrealistic to imagine conformity as the only mechanism at work. Somehow, a more general approach must be formulated.

One might hope, at this point, that social psychologists, while neglecting nationalism per se(22), would have formulated more satisfactory models for the flow of ordinary psychological variables, models which would predict the seven variable-clusters of nationalism as one special case. In the conflict literature, however, the concepts one normally sees from social psychology tend to involve very specific variables, such as frustration, aggression and status inconsistency. A fascinating exception to this generalization is the article by Schwartz on prerevolutionary society, in the Feierabend anthology(23). Schwartz's approach involves the heavy use of approach-avoidance diagrams, with nodes representing clusters of psychological variables and with signs attributed to connecting lines which represent associations between clusters.

Approach-avoidance theory allows him to predict the likely trends in the level of association - plus or minus - between two clusters of variables, and also in the "distance" between them. The same rules for determining these trends could be applied when more objects of thought are brought into the model; also, they might be applied to other phases of social psychology, such as the variables making up nationalism. Beyond its capacity for being generalized, Schwartz's approach has one other virtue: it evokes a detailed picture of the human, psychological feeling of the societies he describes, a picture which he validates in detail from verbal descriptions, yet a picture both sharper and clearer than the usual verbal summaries.

More generally, following up on Schwartz's approach, one might hope to work towards open-ended mathematical models of human behavioral psychology, models capable of achieving greater and greater accuracy as one accounts for more and more variables, within the same mathematical structure, a structure which nonetheless makes substantive predictions. This possibility may be compared with the possibility of predicting weather, by using a set of differential equations rewritten into the form of difference

equations, so that predictions may be made based on knowledge of the initial conditions only at a fixed set of weather stations on a national grid. As one expands the number of weather stations, and makes the grid ever finer, one can make better and better predictions, using the same differential equations as one started with. If one described the "initial conditions" of a human mind in terms of some sort of network structure, and if one's equations specified how to predict the future of any mental network from its present state, then a similar flexibility should be possible in social psychology.

In the limit, as one allows the hypothetical possibility of knowing the initial values of all the psychological variables in someone's mind, one would hope that one's model would approach equivalence to a general cybernetic model of human intelligence and motivation. On the other hand, if one allowed only for very limited knowledge, one would hope that one's social-psychological model would help one choose the aggregate variables of greatest predictive power in making concrete predictions. In the middle-range, one may have to encompass the studies by political scientists such as Sheldon Kravitz(24) on larger-scale psychological structures, as revealed in the voluminous

data of public political statements. Any of these constraints is difficult enough to satisfy by itself; the combination goes well beyond the range of the present discussion.

Given that it may take a long time for anyone to construct such high-level models, and a long time to learn how to deal with them, a lower-level generalization of our simple communications model may have some value as an intermediate step. Instead of starting from a full-fledged model of individual psychology, let us consider a simple equation, drawn essentially from Minsky and Selfridge(25), to describe the changes of a psychological variable, X, under the influence of a "reinforcement" variable, E:

$$X(t+1) = (1 - \theta)X(t) + \theta E(t) \qquad (6.23)$$

In essence, E measures the individual's feeling, after the fact, of what X "should" have been, vis-a-vis his experience at time t. If experiences of E occur at a certain frequency, F, we may approximate this by a differential equation:

$$\frac{dX}{dt} = kF(E-X) \qquad (6.24)$$

This model of individual psychology would leave

out the crucial fact of interaction between different psychological variables; in practice, for example, a strong irrational narcissism on the psychiatric level may provide a pressure towards greater overestimation of the potency of one's nation as well. Again, a thorough description of such interactions would be very complex. However, one may make a simplification based on the idea of cognitive dissonance. If X_e is the value of X one would "expect", or at least find most plausible, based on one's current psychological state with regard to other variables, and if C is the level of confidence with which one feels this expectation, then one might generalize (6.24) to:

$$\frac{dX}{dt} = k_1 F(E-X) + k_2 C(X_e - X). \quad (6.25)$$

In a sense, we have added a new source of "reinforcement" to X, or a new "pressure" on the individual's psychological state.

Finally, it is easy to synthesize this simplified model of individual psychology with (6.22):

$$\frac{dX_i}{dt} = k_1 F_i(E_i - X_i) + k_2 C_i(X_{e,i} - X_i) + k_3 \sum_j C_{ji}(X_j - X_i) \quad (6.26)$$

With the help of a moderately complex model of individual psychology, one might predict C_i and X_{ei} in a very complex way from one's knowledge of the other psychological variables applying to person number i; however, this, by itself, would not require us to change the final coupling terms, on the far right of our equation. Thus the model here can be extended in a fairly straightforward way, building on the work of personal (vs. social) psychologists.

Also, one could develop the model further by learning which measures of C_{ji} are most appropriate, when. For example, one might explore the hypothesis that C_{ji} includes only close family communications, for X_i which represent basic emotional attachments, in communities which have adapted to a combination of intense conflict and extensive ordinary commmunications for centuries. Or one might explore the hypothesis that tacit norms for cooperation depend most heavily on "C_{ji}" measured in terms of constructive bargaining (or other mutual coordination of effort) rather than simple trade or general communications; this hypothesis, if validated, would imply that the reduction of nationalism, and the stabilization of peace, depends critically on efforts by nations to achieve joint benefits from concrete joint activities going beyond

simple trade and ordinary cultural exchanges.

Given that the model above is highly linear, it does not require us to fall back on the even worse approximation of predicting the "average man", as if all people in a given nation were the same; it would allow us to deal with the flow of psychological variables by mathematics quite similar to those well-established for problems such as heat flow. For example, if national narcissism led to uniform values for C and X_e (e.g. a high estimate, X_e, of relative national strength) in a nation, and if there were many levels of communication separating the decision-makers and the people who experience the raw data directly (i.e. realistic E, with a high level of F), one would expect a simple geometric decline in the level of X_e-X with increasing distance of communication; with a deep enough hierarchy, the perceived variable, X, may reflect only the prejudices of the nation, the X_e, and have no reality content at all. Thus one would predict a form of "groupthink"(26), based on large-scale communications effects.

In other cases, however, it may be more appropriate to treat a national communications system as a conglomerate of distinct subsystems (e.g. elite, burghers and masses in nineteenth-century Germany, or

workers and industrialists in modern Japan); this is especially important when one considers domestic conflicts, which may lead in turn to revolution and to messianic nationalism as part of a common pattern. Given a description of a communications system, conglomerate or continuous or a mixture of the two, and given the exogenous data, equation (6.26) would be fairly manageable in providing predictions of the continuous psychological variables of one's choice. Panel survey studies would be possible, to refine the model or provide the data for future predictions, when aggregate national data are inadequate. Classical models of history and of conflict may fit in, by helping us to predict the variables left exogenous in (6.26).

In brief: the concept of communications terms has led us to a generalization - equations (6.21) - of the Deutsch-Solow model of assimilation and political mobilization. In section (v), we will discuss the empirical tests we have given this model in Norway. This generalization of the Deutsch model, while limited in the present context, offers numerous possibilities for important extensions in the future.

(v) ASSIMILATION AND COMMUNICATION: THE CASE OF NORWAY

Broadly speaking, the investigation of variants of equation (6.21), by use of the approaches discussed in Chapter (III), has led to results similar to those of sections (ii) and (iii) of this chapter. The power of communications terms, and of ARMA models, vis-a-vis simpler regression models, has been validated, both in terms of statistical likelihood and in terms of long-term predictive power. The validation has been more significant here, due to the larger quantity of data, but the actual improvements range about 10% in terms of reducing the size of errors. Also, when "outliers" are present, the ARMA models appear much worse in terms of their formal likelihood than do the regression models, though they retain a superior capacity for long-term prediction. Erratic noise, in the form of "ratchet effects," which occur erratically like simple outliers but then <u>persist</u>, does not appear to reduce the modest superiority of the ARMA techniques. In this research, a substantive explanation has also been verified for some of the inconsistent results reported with "gravity models" to predict communications intensities; gravity models were

used to allow us to construct three extra indices of communications within Norway.

In order to test out equations (6.21), statistically, we needed to find a case history with the following characteristics: (i) extensive data on language preference, region by region, distinguishing between urban (mobilized) subregions and rural subregions; (ii) data on the matrix of communications between one region and another region, not aggregated in the form of "total communications entering" or the like; (iii) significant variation across time in a large number of regions in the percentage of language use. Four countries were considered as interesting possible case histories early in this study - Canada, Belgium, Finland and Norway. (Various parts of the British Isles and Africa also seemed promising, but not on the basis of data available at Harvard libraries.) All four have extensive data, commonly available, on language usage. In Canada, however, the data commonly available are aggregated at the level of provinces; except for New Brunswick, most of the provinces of Canada have been consistently close to the extremes of 100% French or no French. In Belgium, the censuses of language were separated by long intervals of time,and the geographical divisions appeared to be just as

sharp, on the whole, as in Canada; the slight variations in Brussels and in Brabant were not enough to change this picture. Of the two remaining countries, Norway had much better data on communications variables.

The data from Norway turned out to be quite good for statistical purposes. Einar Haugen(27), in his book on language problems in Norway, has shown a map, giving the percentage use of the minority language, Nynorsk, in schools in Norway, in the three years 1931, 1945 and 1957, in each of the eighteen provinces of Norway. The data for Oslo (consistent avoidance of Nynorsk) do not appear on the map, but can be reconstructed from the Norwegian Official Statistics which constitute our own source of data; thus we can add in Oslo, to arrive at nineteen major regions in Norway. If we ask how great the gap was, in each region, between the maximum percentage of Nynorsk taught in these three years, and the minimum percentage, we find that these variations across time have been quite substantial. In only six of the nineteen has there been no variation; in the six with the highest variation, the average variation was by 28.5%; in the middle seven, the average was 12 1/7 %. (For example, in Oppland, one of the largest provinces on the map, the percentage use of Nynorsk was

15% in 1931, 44% in 1945, and 40% in 1957, yielding a maximum variation of 29%.) Thus in predicting the variations of language use across time, we are not predicting a dummy variable. Given that the equations in our models all attempt to predict language use at time t+1, controlling for the independent variable of language use at time t, the different averages of language use in different provinces do not water down the effective size of the data sample. In Haugen's map, it also seems clear that the greatest reductions in the use of Nynorsk were concentrated in "intermediate" provinces, or, more generally, provinces which have a significant Nynorsk population but which have a high percentage of communications with non-Nynorsk regions. (Some of the northern provinces, which are far from all the populated parts of Norway, have very strong communications with Oslo, on a relative basis, according to our migration data.) In this study, the use of percentage variables instead of numerical totals helped insure that the results are not dominated by a handful of large subregions.

Data was available in Norway from 36 regions (urban and rural parts of each province, considering Oslo as the urban part of Akershus and the city of Bergen as the urban part of Hordaland), for every year

from 1938 to 1969(28). If we treat all 36 strings of data as sample strings, each with N=31, each generated by the same general process (i.e. governed by the same equations and coefficients), the effective N of the overall sample was 36 X 30 = 1080. (30, not 31, because of the time lags; each string contains 30 pairs of data of language use at time t and language use at time t+1.) Data were also available on: (i) migration from each subregion to each other subregion in the three years 1966, 1967 and 1968(29); (ii) outgoing long-distance telephone calls, total, from 1938 to 1957, by year and by subregion(30); (iii) total letters posted, in 1938-1940 and 1944-1968(31); (iv) births, deaths and marriages from 1938 to 1968(32); (v) real income, from 1938 to 1968(33); (vi) other information on population, crime rate and rate of welfare payments not used in this study(34).

Initially, in coding this data, we were confronted with two interesting choices:(i) whether to define the subregions of each province as "urban" vs. "rural", or to define them as the collections of townships which happened to be defined as urban or rural, in an arbitrary base year, such as 1958; (ii) what to do about the one case of zero data, the case of Finnmark (the northernmost part of the entire mainland of

Scandanavia), where no languages at all were used in schools during World War II, apparently because schools were shut down. In a normal statistical study, one would tend to account for the limitations of multiple regression, and choose definitions for one's variables in order to make them as manageable and as predictable as possible. The moving of townships from the rural to the urban category often had a "ratchet" effect, producing an appearance of change by jerky movements instead of just continuous observable movements. In this case, however, the original concepts of Karl Deutsch clearly called for urban vs. rural percentages, not for geographical subregions; also, it was important to the evaluation of the statistical technique to see if it was as sensitive to ratchet effects - which would appear to be quite common in politics - as multiple regression is; finally, the data on urban vs. rural language use were relatively accessible, while the consistent use of a fixed group of townships would have required approximately fifteen additions and checks for each of 1110 subregion-years, for each of ten variables or so. With Norwegian postal data, an aggregation of this sort was unavoidable, given that the data were available on a township basis but not on an urban vs. rural basis for most years. In the case of Finnmark, we

decided, implicitly, to set the assimilated-percentage variable to zero, in the war years, in the first of our runs, partly as a test of the statistical method. (Finnmark has consistently used no Nynorsk; i.e., it has been 100% assimilated.) This had the effect of introducing a few substantial outliers into the data; this fact turned out to be quite interesting in the runs which followed.

Nine good runs were carried out to predict Norwegian language data, after the prototype version of the "ARMA" program was fully checked out, and the data in the computer checked for consistency with the original data-sheets. The years 1939-1967 were chosen as the main focus of study, to avoid calibration problems with different variables. The first seven runs were carried out on the original data, with outliers existing in Finnmark. By and large, these runs were rather disappointing.

In the first run, ARMA tried out two simple models to predict the percentage, A, of language assimilation in Norway:

$$A(t+1) = bA(t) + c + a(t), \quad (6.27)$$

where a(t) is a random noise term to be minimized, and:

$$A(t+1) = \theta A(t) + a(t) + Pa(t-1), \quad (6.28)$$

where <u>both</u> terms on the right are noise terms, indicating the presence of more complicated noise. Notice, with the regression model, (6.27), we have included a constant term, c, while the ARMA model, (6.28), described in the notation of Chapter (III), does not include a constant term. Thus both models have the same number of coefficients to estimate. With the prototype version of ARMA, the constant term was consistently included in the regression model, and deleted from the ARMA model, to insure that models of the same general level of complexity were being compared.

The results of the first run were relatively disappointing. The regression model received a likelihood score ("LogP", in the notation of Table VI-11) larger than that of the ARMA model, based on the standard normal distribution test described in Chapter (III); the gap in scores was equal to 7, indicating odds of e^7=1100 to 1 against the ARMA model being better than the regression model, empirically. (See section (v) of Chapter (II) for a more thorough discussion of the traditional concepts here.) Given the large data set, this meant that the percentage of variance explained - R^2 - was 99.49% for the regression

model, versus 99.48% for the ARMA model. In a test of long-term prediction, however, the ARMA model did better, as our reasoning in Chapter (V) might have indicated. The regression model had average percentage errors of 14.9%, versus 14.1% for the ARMA errors. Note that we define the "percentage error," in any year, as the gap between the prediction and reality, expressed as a percentage of the averages of the prediction and reality; note also that these errors were averaged by the root-mean-square ("R.M.S.") method. The "absolute errors" in predicting the percentage of assimilation averaged out to 28% for the ARMA model, and 39% for the regression model; the huge figures are due to occasional wild predictions, building up geometrically from 1939 to 1968.

Note that a 2% reduction in square error, from 1 - .9948 to 1 - .9949, is considered highly confirmed by the usual likelihood test, with a sample this large; the larger reductions in long-term prediction errors by the ARMA routine would appear to be even more certain in their validity. Indeed, it seems much more suspicious in some ways to discuss the reduction of very small errors - about .50% - than to discuss the reduction of more substantial errors. In theory, the classical likelihood measure is enough to account for

such "multicollinearity," but even so such situations have often turned out to cause problems for statisticians. Note that the presence of multicollinearity would presumably be much worse than here, for processes which one would hope to find more predictable in the long-term; the inability of classical approaches to perform well under such circumstances is one more reason to favor a new approach. An approach which attempts directly to minimize the more substantial errors in long-term prediction would appear to be much safer. Also, as in section (iii) of this chapter, it is critical that formal statistical likelihood and predictive power have not gone hand-in-hand in their evaluations of the different models available.

In later runs, we hoped that the ARMA models would do better in terms of statistical likelihood. After all, the constant term in the regression model could reflect a trend away from Nynorsk, a trend which could be explained by communications terms and other terms, so that the value of a constant term as a surrogate variable would disappear when they are accounted for. Also, for reasons described in sections (v) and (vi) of Chapter (II), we hoped that the ARMA model would be more sensitive to terms of realistic importance,

increasing in likelihood by more than the regression model does when such terms are included.

In the second run, we decided to introduce a communications term. In Norway, we had only one explicit measure of communications from each province to each other province available - average migration from 1966 to 1968. Given that population data were not fully available from 1939 to 1967, and given that intraprovincial communications are presumably not well-measured by internal migration data directly, we used the following simplified model:

$$A_i(t+1) = c_1 A_i(t) + c_2 \sum_j M_{ji}(A_j(t) - D_j(t)) + c_3(S_i(A_i(t)-D_i(t))+S_i^*(A_i^*(t)-D_i^*(t)))S_i, \quad (6.29)$$

where M_{ji} represents migration from region number j to region i, where S_i represents the sum over j of M_{ji}, and where the asterisk refers to variables in the region complementary to region i. (i.e. A_i^* is the percentage of assimilation, measured in the same province as region i, but in the rural part, if region i is urban, or in the urban part, if region i is rural.) In principle, M_{ji} should have been divided by the population of region i, but this was not only

impossible, it was of limited potential importance in a nation with provinces of comparable population. In retrospect, it might have been better to start out with the full, more complex model, (6.21), even in these early investigations; however, due to the difficulties and potential controversy in estimating the shape of $f(D_i)$, it was decided that priority should be given to the simpler formulation at this stage.

At any rate, the model written out above - (6.29) - did not perform especially well, with our initial Norway data, according to the usual statistical tests based on short-term prediction. In terms of statistical likelihood, the regression model in this run did no better than the regression model of our earlier run, without communications terms. (Gaps in likelihood less than one point, as discussed at the base of Table VI-11, were not recorded, due to the implication of no significance in such differences in apparent performance.) In other words, the extra terms did not appear to add anything. The estimate of "c_3" here, as in all the other runs carried out on this model and its analogues, was too small for the computer output formats to cope with. "c_2", however, was on the order of 1%. Again, the regression model was superior to the ARMA model, with a gap of likelihood scores of

5.65, indicating odds of 280 to 1 favoring the regression model. The ARMA communications model had a slightly higher likelihood - by 1.5 - than the simple model (6.28), indicating odds of 4.5 to 1 in its favor; however, these are not exactly overwhelming odds. The R^2 of the ARMA model was .9948, versus .9949 for the regression model, just as before.

In long-term prediction, however, from 1939 to 1967, the ARMA model did increase its margin of superiority; its R.M.S. average percentage errors were 13.4%, versus 14.7% for regression, while its absolute errors were 27% versus 39%. The communications term, even if poorly estimated, clearly added something to longer-term prediction. With a different estimation approach, oriented towards predictive power instead of maximum likelihood, the gain provided by the communication terms might have been considerably larger. Also, with the original model, (6.21), the intermediate provinces of Norway, instead of the minimum-Nynorsk provinces, might have been singled out more effectively as likely areas of large-scale assimilation; again, the predictive power of the model might have been enhanced.

In the third run, as an alternative hypothesis, we considered the possibility of using real income as a

variable to predict language:

$$A(t+1) = c_1 A(t) + c_2 Y(t), \qquad (6.30)$$

where Y(t) represents the real income in a region, and where we have not written out the noise terms. In terms of likelihood theory, the performance of this model was exactly the same as that of (6.29), as described above. In long-term prediction, however, it did not do quite as well. The ARMA R.M.S. average percentage errors were 14.1%, and absolute errors 28%; the regression percentage errors were 14.9%, absolute errors 39%. These results are closer to those of the univariate models, (6.27) and (6.28), in quality, than to the results with (6.29). In principle, however, the relative potential of the two models in long-term prediction will not be clear until a new type of estimation system is available.

In the remaining runs on our original data, we decided to explore communications indices other than that of simple migration. Two other measures of communication - telephone calls and volume of mail - were available; however, these were only available on a province-by-province basis. We faced the problem of how to reconstruct the matrix of communications from each province to each other province. This problem has

often been faced elsewhere in regional science(35) and in sociology, and resolved by way of a "gravity" model. The original gravity model, proposed by Stewart, would estimate province-to-province telephone communications, for example, as follows:

$$C_{ij} = c_0 \frac{T_i T_j}{r_{ij}}, \qquad (6.31)$$

where T_i and T_j are the total volumes of telephone communication (or <u>other</u> communications variables, such as migration) in each province, and where r_{ij} is the distance between provinces. A modified version, studied by Galle and Taueber(36), and discovered to have a multiple correlation of between 89% and 93% between prediction and reality, is as follows:

$$C_{ij} = c_0 \frac{T_i T_j}{r_{ij}^k}, \qquad (6.32)$$

where k is an unknown exponent to be estimated. (Note that the correlation here, with a <u>cross-sectional</u> model, is stronger in its implications than a 90% would be in a predictive time-series model, <u>insofar</u> as a cross-sectional study makes sense here.) Curiously enough, while equation (6.32) has been successful in empirical tests, the parameter "k" has varied a great

deal in its estimated value; Galle and Taueber report that k was equal to .62 for interurban migration in the US in 1935-1940, but only .42 for the same data as measured in 1955-1960.

In order to estimate the likeliest value of "k" for interregional communications in the case of Norway, it was necessary to fit the model (6.32) to the only region-to-region communications data available - again, the migration data. A direct fit of (6.32) would have required the use of nonlinear regression; however, equation (6.32) can be transformed as follows:

$$\log C_{ij} - \log T_i - \log T_j = a - k \log r_{ij}, \qquad (6.33)$$

where the entire left side of the equation forms the dependent variable, and where "a" and "k" can be estimated by multiple regression.

As long as we were carrying out such a regression, however, it seemed appropriate to test out a new explanation for the reduction in "k" from .62 to .42 as measured in the United States by Galle and Taueber. It is fundamental to the communications theory of nationalism, as described in section (iv), that there has been a historic rise in the strength of long-distance communications, <u>relative</u> to

shorter-distance communications, at least for the average man. Using equation (6.32), we may compute the ratio of communication across a long distance, R_1, to the communications across a shorter distance, R_0, between regions of equal size (T_i the same for all regions):

$$\frac{C_1}{C_0} = \frac{\frac{1}{(R_1)^k}}{\frac{1}{(R_0)^k}} = \left(\frac{R_0}{R_1}\right)^k \qquad (6.34)$$

For given distances, R_0 and R_1, the terms involving c_0 and T_i, etc., cancel out; thus the only way this ratio can get larger, for a given comparison of distances, is if k gets smaller. (e.g. A small variable to the zeroth power will equal 1, which is the maximum this ratio can approach under the stated conditions.) Thus it is critical to our communications theory that k should tend to decrease in time, as the result of some aspect of "modernization"; the most obvious aspect of "modernization" to consider is the economic factor, the increasing income of people relative to the cost of communication. Thus we decided to test the model:

$$k(t) = c_1 - c_2 Y(t), \qquad (6.35)$$

where Y represents the real income of a region. Also,we decided to consider the possibility that a term representing social proximity (urban vs. rural similarity of regions) should be accounted for. Thus, in the final regression equation, we decided to test:

$$\log M'_{ij} = \log M_{ij} - \log S_i - \log S_j$$
$$= c_0 - c_1 \log r_{ij} + c_2 Y \log r_{ij} + c_3 U_{ij}, \quad (6.36)$$

where U_{ij} is defined to equal one if both regions are rural or both urban, but zero if they differ; a measure of distance was obtained from the World Atlas(37); S_i is defined as with (6.29). Note that Y(t) - real income in the subregion from which migration occurs - was not a surrogate for time in this regression, since the regression was based on a combination of two 36 X 36 matrices of total migration in the close-by years 1967 and 1968; the primary variation in real income was between subregions. The covariance matrix produced is

shown in Table VI-25:

	U_{ij}	Y log r_{ij}	log r_{ij}	log M'_{ij}
U_{ij}	.251	.017	-.012	.072
Y log r_{ij}	.017	1.299	.286	.135
log r_{ij}	-.012	.286	.875	-.549
log M'_{ij}	.072	.135	-.549	1.062

Table VI-25: Gravity Model Correlations

Inverting the three-by-three matrix in the upper left of this table, and multiplying the inverse by the vector formed by the three upper numbers of the rightmost column, we can compute the standard regression coefficients for (6.36):

$$c_3 = .24$$

$$c_2 = .26$$

$$c_1 = .70,$$

all with the expected signs, and all clearly very significant for the large N we have considered and for the variances displayed in Table VI-25. Thus our income hypothesis appears to have been validated rather strongly. This same regression analysis was also used to construct an approximate measure of communications for one of the other communications variables available in Norway, telephone communications; however, c_2 was

deleted from the regression equation, and c_1 thereby reduced to .67, due to the computational difficulty of calculating the full index, as based on different values of the income variable across time.

In the fourth run on Norwegian language data, equation (6.31) was used to reconstruct the matrix of telephone communications, C_{ij}, to replace M_{ij} in equation (6.29). The years 1939 to 1957 were used as a data-base. Once again, the regression model did better than the ARMA model in terms of log likelihood, with a gap of 3 points, implying odds of 20 to 1 in favor of the regression model. The R^2 of the regression model was .9941, versus .9940 for the ARMA model; this was substantially worse than our earlier runs. On the other hand, this was substantially worse than our earlier univariate runs, encompassing a subset of the independent variables here; this signals us that the data in the period 1939 to 1957 average out to be more difficult to predict than the previous data-base, 1939 to 1967. Indeed, these years contain all of the wartime "outliers" mentioned above, in Finnmark. In light of these difficulties, the model did relatively well in long-term prediction. The R.M.S. average percentage errors were 13.7% and 14.1% for the ARMA and regression models, respectively; the R.M.S. average absolute

errors were 35% and 40% for the two models, respectively. Perhaps a later run without outliers would have given a much better picture.

In the fifth run on Norwegian language data, we studied the same model as in the fourth run; however, this time we used equation (6.36), adapted to predict telephone communications, to reconstruct a matrix of telephone communications. The R^2 and the likelihood scores turned out to be the same as in the fourth run, except that the ARMA model gained very slightly in likelihood - by one point; the odds against this being a coincidence are only 3 to 1, according to likelihood theory - not a substantial confirmation. The results of this run seemed sufficiently bad, with R^2 still low, that simulations were not carried out.

In the next run on Norwegian language data, postal data were used, with equation (6.36), to construct an index of communications, to replace M_{ij} in equation (6.29). The regression model performed better than the ARMA model, with a gap in likelihood of 9, implying odds of $e^9 = 8100$ to 1 in favor of regression. The R^2 of the ARMA model was .9958, versus .9959 for the regresson model. At first, these high values of R^2 seemed rather encouraging. However, the data period used for this analysis was 1945 to 1967, due to the

absence of postal data in three of the war years; this implied that the outliers in Finnmark were avoided. In our seventh run, as a corrective, we re-evaluated the simple univariate model, equations (6.27) and (6.28), over the same time-period; the results were the same as with the postal model - implying that nothing was gained by adding these communications terms - except for an insignificant one-point decrease in the likelihood of the ARMA model.

After these seven runs were completed, a careful review was carried out, first of the ARMA models, and then of the communications models. Another run was carried out on a different set of data - on births, deaths and marriages as a single set of variables. In that case, the ARMA model outperformed the regression model by 339 points, by the usual likelihood measures, implying an astronomically high probability of its superiority. In conventional language, this gap of 339 points implies that,"the ARMA model was confirmed with a p less than 10^{-100}." Concretely, the ARMA model had an R^2 of .975 in predicting the marriage rate, versus .85 for regression; also, the variance of the errors in predicting the death rate was reduced by 10%. Unfortunately, the computer refused to calculate a full table of predictions for this case, because the table

was too long; however, the ARMA model did not seem notably superior to the regression model in that portion of the predictions which the computer did print out.

In order to explain the mediocre performance of the communications models, we looked very closely, province by province, at the direction and shape of the errors made by the ARMA communication models (from run number 2) in long-term prediction. There did not seem to be a notable tendency for errors to be biased in one direction in any special group of provinces, except for the "intermediate province" group which the communications terms should have been able to distinguish; however, there did seem to be a very strong tendency for the predictions to be systematically low, everywhere. With constant terms, of course, this would not have been expected. Still, the independent variables in equation (6.29) were close enough to being able to represent constant trends, upwards, that it seemed very strange that such a bias would develop. A series of intuitive arguments convinced us that the outliers in Finnmark, extending for a handful of years in both urban and rural Finnmark, could add a degree of apparent randomness, enough to bias the coefficients substantially. Given

that Nynorsk had never been used at all in schools in Finnmark, we felt it would be best to change the data, so that in all years the variable "A" (percentage assimilated) would equal 100% in Finnmark.

After these changes, two runs were carried out, both highly successful. The first run duplicated our original first run, based on equations (6.27) and (6.28). This time, the R^2 for the ARMA model was .9988, versus .9987 for the regression model. The ARMA model had a likelihood score of 35.46 points higher than the regression model, implying odds of 2.5 million billion to 1 against its superiority being a coincidence. Both models, of course, were doing astronomically better than any of the models discussed above. In simulation, however, the picture was a bit mixed, though still improved on the whole. The R.M.S. average percentage errors were 9.9% for the ARMA model and 9.4% for the regression model; the absolute errors averaged to 20% and 27%, respectively.

The second new run duplicated the second old run, in using equation (6.29) as a model. The superiority of the ARMA model grew larger, when a more complete substantive model was used, just as we had hoped earlier; the gap in likelihood grew to 42.51, implying odds of 3×10^{19} to 1 in favor of the ARMA model. With

the regression models, the addition of communications terms produced only a slight gain in likelihood - 2 points - but with the ARMA model, the gain in likelihood was 9.03 points, implying a very significant improvement. (Significant at the level of "p = .00011," in conventional terminology.) One may note that the coefficient of the main communications term was .0067 for the ARMA model, versus .0054 for the regression model, both about right for a recurrent feedback term. With a better substantive model, the ARMA model improved much more in its long-term predictive power, too, than the regression model did. The average percentage errors were 8.6% for the ARMA model, versus 9.0% for regression; the absolute errors averaged to 17% for the ARMA model, versus 25% for regression. Between the two measures, it is reasonable to say that the ARMA model here, as elsewhere, displays on the order of 10-15% less error in long-term prediction than the regression model does. Also, as we pointed out at the beginning of this section, when outliers are removed, the ARMA models are very much superior to the regression models in terms of formal statistical likelikood. These statements remain true despite the "ratchet" effects - similar to outliers, but persistent - which we mentioned earlier in this

section.

FOOTNOTES TO CHAPTER (VI)

(1) Deutsch, Karl W., Nationalism and Social Communications, MIT Press, Cambridge, Mass., 1966, revised second edition. Chapter 6 contains the main argument leading up to the mathematical model; Appendix V contains the mathematical model, and a verbal description of it.

(2) ibid.

(3) Hopkins, Raymond, "Projections of Population Change by Mobilization and Assimilation," Behavioral Science, 1972, p.254. The programs were made available to us by Prof. Deutsch at the Harvard Department of Government.

(4) Deutsch, Karl W., op. cit., Appendix V. Note that several versions of this model have appeared in print. The version here, in all fairness, was actually taken directly from Hopkins, Raymond and Carol, "A Difference Equation Model for Mobilization and Assimilation Processes", 1969, unpublished; a copy of this paper was provided to us by Prof. Deutsch, and described by him as containing the final revision of the model. This revision appears, in difference equation form, in Hopkins, Raymond, "Projections of Population Change by Mobilization and Assimilation", Behavioral Science, 1972, p.254. The reasons for the revisions to earlier versions are described in Hopkins, Raymond, "Mathematical Modelling of Mobilization and Assimilation Processes", in Mathematical Approaches to Politics, edited by Hayward Alker, Karl Deutsch and Antione Stoetzel, Elsevier Publishing Co., New York, 1973, p.381.

(5) See note 1.

(6) Hopkins, Raymond, "Mathematical Modelling of Mobilization and Assimilation Processes", in Mathematical Approaches to Politics, edited by Hayward Alker, Karl Deutsch and Antione Stoetzel, Elsevier Publishing Co., New York, 1973, especially p.381.

(7) See note 3. More precisely, we used the Hopkins routines directly, on sample cases suggested to us by Prof. Hopkins and on a few others.

(8) In Chapter (V), we noted that the correlation across n intervals of time, when both process noise and measurement noise might be present, would equal $\phi^2 r^n$, where ϕ is the correlation involved with measurement noise, and r is the true correlation across time of the process underneath. When this figure changes very little with increases in "n", but is substantially different from 1 for various values of n, then it would seem that r is very close to 1 and that ϕ is not.

(9) The approximation here is that exp(1+a) is approximately equal to 1+a, with only about .5% error when "a" is about 10%. More precisely, if we take the exponential function of both sides of the equations (6.10), and substitute in from (6.12), the approximation rule cited here brings us back to (6.9).

(10) Strictly speaking, there is one major qualification one might make to this statement. When making a prediction, one usually starts from a given base year, and applies the differential equations to that year as an initial condition. This would correspond to <u>adjusting</u> k_7 and k_8 here, to fit a given year exactly. One can expect to do better, if one somehow averages different base years to get an estimate of the underlying reality, and uses that estimate to make predictions from. Admittedly, part of the advantage in our "ext1"(EXTRAP) extrapolation probably lies in doing just that. If there were a consistent change in the rate of growth of these variables, through time, and if one were using extrapolation models to predict the same period of time as the one they were fitted to, this would lead to an <u>unfair</u> advantage for the extrapolation models; the extrapolation models would be centered at the middle of the process, but the ordinary models would be centered at the initial extreme. However, every one of the extrapolation runs here was accompanied by a run testing the predictive power of the hypothesis of a t-squared term in (6.10); these runs gave no support to the idea that factors involving a simple second derivative could be responsible for the advantages of extrapolation. The ability of extrapolation to average out extreme values measured in the <u>same</u>, early periods of time is not "unfair," insofar as it reflects an advantage available to those trying

to predict the future from an extensive data-bank from the present and past.

(11) The computer printout here was left in the custody of Prof. Karl Deutsch, Harvard Dept. of Government. The computer printouts from section (ii) were also left in his custody, in 1971. These two groups of output are approximately one foot thick, put together. For every run reported here, they include predictions and reality for every year, past and future, for which predictions were made.

(12) Deutsch, Karl W., op. cit., Chapter 6.

(13) Lieberson, Stanley, Language and Ethnic Relations in Canada, Wiley, New York, 1970. See p.47, 48, 183-187 for relatively smooth graphs emerging from scatter-plots.

(14) ibid. Lieberson focuses strongly on the issues of language "retention," by those brought up in one language, as opposed to "demographic factors." On p.35, he states that,"It is far more correct to describe the Canadian scene as an equilibrium based on counterbalancing forces." On p.50 and p.51 he emphasizes, first, that English has been dominant in terms of "retention" or assimilation, but then, that the "revenge of the cradle" has been central to French language maintenance. On p.225, he defines a variable, "communications advantage," quite similar in spirit to the "language pressure in communications" discussed here; in the subsequent verbal discussion, he implies that this variable is central to "retention" phenomena.

(15) Deutsch, Karl W., "Mathematics of the Tower of Babel", in "Nation and World", in Contemporary Political Science: Toward Empirical Theory, McGraw-Hill, 1967.

(16) Strictly speaking, one must also try to explain the origins and convergences of such dialects, instead of merely the decision by individuals to jump from one dialect to another. Equation (6.26), which can deal with the idea of dialects getting closer or further away as a result of communication, is conceptually quite close to the model here. Yet one is still faced with the

problem of explaining how divergence in dialect can come about. If the speech in each region were subject to random drift, or to systematic pressures based on interregional differences in speech equipment, then a low level of interregional communications, in our model, would imply little damping of such drift. An increase in communications would imply a greater pressure for convergence, and a damping out of future drift. Note that such phenomena would also apply if there were a dramatic increase in communications between two regions with enough internal communications to resist assimilation as such; the growth of "franglais" is an interesting example.

(17) Schelling, Thomas C., The Strategy of Conflict, Oxford U. Press, New York, 1963. (Copyright 1960.) p.104: "But where do the patterns (of potential compromise) come from? They are not very visibly provided by the mathematical structure of the game, particularly since we have purposely made each player's value system too uncertain to the other to make considerations of symmetry, equality, and so forth, of any great help. (i.e. of help in analyzing Schelling's paradigms for games of mixed conflict and common interest.) Presumably, they find their patterns in such things as natural boundaries, familiar political groupings, the characteristics of states that might enter their value systems, gestalt psychology, and any cliches or traditions that they can work out for themselves in the process of play..." p.151: "... the introduction of uninhibited speech may not greatly alter the character of the game, even though the particular outcome is different..." In short, tacit norms, before the introduction of explicit bargaining, are crucial to the existence of possible "patterns of convergence." On p.113-114, Schelling hammers home the point that mathematical "solutions" to nonzerosum games do not provide a realistic alternative to his own theory of tacit norms, discussed on p.99-111. In a sense, one might argue that the idea of "solving" for a unique or optimal static equilibrium may apply only to games similar to those originally discussed in such terms by Von Neumann and Morgenstern (note 39 of Chapter (V)); as in economics, there may be situations where dynamic factors cannot be easily encompassed within such a static description. At any rate,

Schelling's discussion, applied elsewhere by him to real political analysis, strikes us as fairly convincing.

(18) Deutsch, Karl W., Nationalism and Social Communications, MIT Press, Cambridge, Mass., 1966, revised second edition, especially p.96-97. Also, Deutsch et al, "Political Community and the North Atlantic Area", in International Political Communities, Anchor Books, New York, 1966, p.17. In the latter reference in particular, the concept of "community" is defined in terms of an ongoing ability to communicate and respond in decision-making processes of a continuous sort.

(19) See note 17.

(20) See note 18, particularly the second reference.

(21) See note 16.

(22) Deutsch, Karl W., Nationalism and Social Communications, MIT Press, Cambridge, Mass., 1966, revised second edition, p.26.

(23) Feierabend, Ivo K. and Rosalind L., and Gurr, Ted R., eds., Anger, Violence and Politics: Theories and Research, Prentice-Hall, Englewood Cliffs, N.J., 1972.

(24) Kravitz, Sheldon, A Theoretical Model For the Analysis and Comparison of Ideologies, Ph.D. dissertation, May 1972. Available c/o Widener Library, Harvard U., Cambridge, Mass.

(25) Minsky, Marvin and Selfridge, Oliver G., "Learning in Random Nets", in Information Theory, Fourth London Symposium published by Butterworths, 88 Kingsway, London W.C.2., U.K., p.339. In discussing this formula, Minsky and Selfridge consider only the cases E = 1 or E = 0 per episode, but the generalization does not appear very difficult. These authors, in turn, refer to Bush, R.R. and Mosteller, F., Stochastic Models for Learning, Wiley, New York, 1955, as a basic source.

(26) "Groupthink" as a small-group phenomenon has been widely discussed as a result of Janis, Irving, Victims of Groupthink, Houghton-Mifflin, New York, 1973.

(27) Haugen, Einar, Language Conflict and Language Planning: The Case of Modern Norwegian, Harvard U. Press, Cambridge, Mass., 1966. Map on p.229.

(28) The primary source for this data, as with the data reported below, was the Norwegian Official Statistics series, commonly available in U.S. libraries. The Norwegian name is "Norges Offisielle Statistikk"; when author designations are required, the "Central Bureau of Statistics" or "Sentral..." is usually appropriate. School data for Jan. 1970 may be found in the 1972 "Arbok" (Yearbook), which also appears as Rekke-XII, number 274. Data on the use of languages in elementary education were used, as on p.335 of that copy of the Arbok. The language use data for earlier years were taken from the earlier Arboks, back to 1939. In some years, when the urban/rural breakdown was not available, we used the Skolestatistik issues of the N.O.S. Every number(issue) in the N.O.S. series includes a list of the numbers and topics of other recent issues; also, on the front or back cover are listed the numbers of previous issues on the same topic. (Thus, in the 1972 Arbok are listed the Rekke and number of all previous Arboks.) Language use in elementary schools is essentially a matter of local choice; Haugen, op. cit., gives a few details of the process of language choice. We decided, in our computer runs, to recalibrate the time periods of the data; thus, language use in force in schools in January 1950 was taken to be an index of actual language use in 1949, given the lags involved in changing policy in the schools.

(29) Sources: N.O.S., op. cit., Rekke XII, No. 233; Rekke A, No. 244; Rekke A, No. 292. Original statistics were further broken down by sex, but aggregated for this study.

(30) N.O.S., op. cit., Rekke XI, No. 298 for the most recent data, and previous items in the same topic series. Note that Rekke XII, No. 232, while not containing the appropriate breakdowns by urban and rural, does provide definitions in English.

(31) N.O.S., backwards from Rekke XII, No. 198. Aggregated according to the urban/rural definitions of townships spelled out in the Skolestatistikk series.

(32) N.O.S., op. cit., backwards from Rekke XII, No. 220. Some aggregation required in earlier years over sex, etc. In 1967, Rekke XII, No. 244 was used.

(33) N.O.S., op. cit. For income, Rekke A, No. 363, used for 1968 on back. ("Municipal Total Income," from Table IA. In early years, some aggregation required; however, the column names were uniform enough that it was not too difficult to reconstruct the same aggregations as used by the N.O.S. authors in later years.) In Rekke XII, Nos. 245 and 252, an index of consumer prices was found, for the entire period. (i.e. An historical table was available.) The average size of the ratio, normalized to units appropriate for statistics, was on the order of unity.

(34) These data include data on criminal convictions - easily available in this period, starting back from the Arbok; data on heads of households on welfare, continued back in the Statistical Monthly over the entire period; data on population, not generally available in recent years with the desired breakdown, but in the Arbok when available.

(35) Isard, Walter, Methods of Regional Analysis: An Introduction to Regional Science, MIT Press, Cambridge, Mass., 1966, Chapter 11. On p.500, reference is made to Stewart and Zipf, the two fathers of the idea; on p.506, a concept of social distance is mentioned, similar in spirit to "U¿j"; on p.507-510, empirical results are discussed. See also Deutsch, Karl W. and Isard, Walter, "Toward a Generalized Concept of Distance", Behavioral Science, Nov. 1961.

(36) Galle, Owen R. and Taueber, Karl E., "Metropolitan Migration and Intervening Opportunities", American Sociological Review, No. 31, Feb. 1966, table on p.8. Note that these authors are essentially critics of the gravity model; thus their results are particularly interesting. For other work in this area, see note 35.

(37) World Atlas, Moscow, 1965, p.57. A large map of Norway, with major roads indicated, was used. Distance was measured with a centimeter ruler, for the most direct major route by road; however, if

this should exceed the absolute distance by 40% or more, then the direct distance plus 40% was used. For distances from an urban area, either there was one major city, or several which could be averaged. For rural distances, it was assumed that population density was even throughout each region; averages were estimated on that basis. All of the data here was punched on cards, and read into the MIT Multics machine; the punched cards and code sheets may be made available to future users through the office of Prof. Deutsch, if there is interest in so doing.

(V) GENERAL APPLICATIONS OF THESE IDEAS: PRACTICAL HAZARDS AND NEW POSSIBILITIES

(i) INTRODUCTION AND SUMMARY

Fierce debates continue to rage between those who would study behavior "with mathematics" and those who would study it "by traditional means." These debates, by drawing attention to the extremes, have obscured many of the serious hazards and many of the most important applications of mathematical approaches in government and in psychology. In extending the mathematical approaches further, we have a special responsibility to discuss the new applications and the continuing hazards which may result.

We will begin, in section (ii), by presenting the viewpoint of the practical decision-maker, who has not used mathematical methods so far, for good reasons. All of this chapter will be organized around the difficulties which he faces; other possible users of our ideas - the social scientist, the psychologist and the ecologist - will be mentioned within more limited contexts. In section (iii), we will suggest a common framework for evaluating verbal and mathematical tools, both, based upon the common goal of prediction; within

this framework, the mathematical methods have a role to play, at least in principle, both for what they tell us directly and for what they tell us about common abuses of verbal methods. For those who prefer to deal in concrete examples, rather than abstract generalizations about methodology, we have included a number of relevant examples, mostly in the footnotes. Also, at the end of this section, we will describe in detail how this framework has motivated the development of new mathematical procedures described in the other chapters of this thesis.

In section (iv), we will go from principle to practice; we will discuss specific ways in which statistical methods may be used, <u>in close relation</u> with verbal methods, and be of significant value in real prediction efforts. This discussion will not be based upon the well-known philosophy of logical positivism, but on the more recent philosophy of Bayesian utilitarianism (see section (v) of Chapter (II)), the philosophy which underlies the actual mathematical developments we have discussed; at any rate, the utilitarian approach helps keep us focused on the value of our methods to serious policy-makers. In this section, we have also tried to crystallize out our own experience with the numerous ways in which statistical

research can turn out to be useless and misleading for the policy-maker, if it is done in a cavalier manner; our suggestions are not a definitive answer to all of these difficulties, but at least they may help.

Finally, in section (v), we point out that the central role of human psychology in politics seriously limits the possibilities of naive empiricism, both verbal and mathematical. Statistical studies, like verbal research of the purely empirical variety, may be unable to transcend these limits. However, the mathematical ideas discussed in Chapter (II), along with other offshoots of the Bayesian approach, can be applied in a different way, to help overcome these limitations in a way which words alone cannot; to illustrate this point, we will mention specific possibilities for using these ideas in the future to cope with and explain the phenomenon of intelligence, whether in human societies or in human brains.

(ii) THE LIABILITIES OF MATHEMATICAL METHODS IN PRACTICAL DECISION-MAKING

Let us start out by reconsidering our tacit assumption that methematical methods do have some use, after all, in political science and in political

decision-making. Many people have questioned this idea in the past few decades, and many more may have felt strong private reservations about the idea. While we clearly can be expected to reaffirm the value of mathematical methods, on the whole, we also believe that the traditional complaints against mathematical methods do contain real information which the user of such methods should not ignore.

In particular, let us try to express the reservations about mathematical methods which might be held by the active political actor. Practicing diplomats and politicians have often found that their margin of success depends on their ability to seize upon unique twists in the political or psychological environment, twists which allow the individual to escape the seemingly uncontrollable tide of events that one would expect a mathematical model to extrapolate. Sometimes this involves the ability to establish channels of serious communications between different political groups, channels which can grow in importance once they have been established. Sometimes this involves the ability to seize upon an economic or military advantage. Caesar's Gallic Wars are a classic example of the latter sort of imagination, evading Lanchester's Laws at every turn(1); Liddell Hart, in

his classic text on military strategy(2), has emphasized that such imaginative approaches have been decisive in wars throughout history. In both cases, one achieves a greater "benefit" within a given "cost constraint", not by being tight and precise about budgeting one's resources, but rather by preserving the detachment and the freedom one will need in order to seize upon whole new options, which may open up a whole new frontier of possibilities. Political creativity in this form is difficult enough to encompass within any scholastic context, let alone the context of mathematical models; therefore, political scientists who have a strong attachment to this process would naturally tend to be skeptical of mathematical models. More generally, successful political actors, like most successful professionals, would tend to believe that they stretch their minds to the limit, in order to arrive at their policy decisions; they may conclude that the sheer complexity of their own decision-making militates against the prediction of its outcome by mathematical systems which account for far less information content. Furthermore, it is also likely that a large part of this information, even when accessible to the political scientist, may be encoded in a verbal form which militates against its being

accounted for by mathematical models.

(iii) PREDICTION: A COMMON GOAL FOR VERBAL AND MATHEMATICAL SOCIAL SCIENCE

These difficulties can be dealt with on several different levels. Let us begin on the simplest level.

The difficulties above point to the impossibility of constructing mathematical models which will predict exactly what will happen in politics, in detail, in the short-term and in the medium-term. However, these difficulties have also been enough to make it impossible for any human being, political actor or otherwise, to predict exactly what will happen in all of politics, in the short-term or medium-term. A traditional political scientist or mathematician might deduce at this point that "true prediction", in the sense of exact prediction, is impossible in political science. Therefore, in order to assure himself that he is involved in serious work, he may restrict his attention to propositions which meet an Aristotelian test of "truth", such as statements about historical documents(3) or abstract theorems which he can prove by rigorous deduction. Very few political actors, however, feel that they would want to turn away from

the difficult but primary question of predicting the differences in outcome between the different actions they could take. These predictions may always include factors of uncertainty, but the political actors would find it interesting enough to reduce this uncertainty as much as possible, in any possible way. Thus, insofar as political scientists are concerned with developing objective insights of the maximum possible value to their consumers, the political decision-makers at all levels, their ultimate concern would be with the development of effective probabilistic theories to predict political and social systems.

There are five points worth noticing about our emphasis on prediction here. First of all, this emphasis does not restrict itself to the overtly mathematical phases of political science. The development of "predictive models" - mathematical or verbal, or even analogue for that matter - is a general concept, which can be used to guide historical research as easily as it guides statistics. Many of the "grand theories" of political history, including especially the theories of Spengler(4), Toynbee(5), Turner(6), Hegel(7), and Marx, were designed to help people "understand" history in terms of a verbal dynamic model which could also be used to predict the future.

Traditional approaches to research in political science, however, might tend to "develop" such theories by adding complex strings of qualifications, and by forcing an elaborate, perfect Aristotelian fit of the weaker, more specialized propositions which emerge. Our own approach would ask that political scientists continually return to the main question, to the ability of their theories, with the disclaimers removed, to predict the broad first-order trends in the major, most obvious variables of political history.

Second, our emphasis on prediction can be justified on deeper grounds than those of satisfying those who pay for the bulk of the political research. Following the philosophy of utilitarianism, one might simply regard political science itself as one particular phase of political activity; one might even suggest that its major justification for existence, in the long term, is its ability to contribute to constructive political activity. This takes us back to the primary need of the decision-maker to predict the results of his actions, at least on a probabilistic basis. On the other hand, even if one were willing to accept the ethical principle that truth should be pursued for its own sake, as an ultimate goal equal to or higher than the goal that of human welfare, one

still faces the problem of defining what this "ultimate truth" would consist of. One might suggest that ultimate truth, if it does exist in political science, lies not in the changeable facts of current happenstance, but rather in the less changeable dynamic laws which lead from one set of circumstances to another; in physics, for example, the dynamic field equations are considered the highest scientific truth, while the codification of the wave-function of the universe is not an object of serious study. Admittedly, our knowledge of the dynamic laws, unlike the laws themselves, is likely to be changeable for a long time, in political science as in physics; however, it would be meaningless to speak about the advancement of knowledge as a worthwhile goal, were there not such a possibility for change and expansion in the state of knowledge. There are those who would question, in varying degrees, the primacy of the most abstract dynamical equations even in a field like physics; however, even the "phenomenological approach", in that field, involves the construction of powerful, generalized predictive statements, statements about what to expect after setting up experiments of different types(8).

Many times in political science, the concepts of

"causation" and "explanation" have been cited as forms of truth worth pursuing(9). The statement that "A caused B" may be translated roughly into the statement that, "A occurred before B, and the dynamics of the system were such that B would not have occurred if A had not occurred when it did, ceteris paribus." Once again, the critical question to answer is that of the dynamic laws which govern political systems.

Beyond the goals of social utility and "ultimate truth", the political scientist might also pursue the goals of cultural enrichment and entertainment. These goals are often cited as a justification for extreme traditionalism in political science. Whether these goals are now being pursued effectively by all of those who cite them is a difficult matter to judge, well beyond the range of the present discussion. However, a large part of the "cultural enrichment" involved would appear to involve the learning of lessons about human psychology, about what patterns of thought and behavior one might <u>predict</u> on the part of human beings or human groups in unusual circumstances, in other cultures.

Third of all, one should note that our emphasis on prediction as the ultimate goal of political science does not imply that work of a more descriptive nature

should simply be abandoned. Using statistical methods, for example, one must first collect a set of data, before one can fit or test a model. After one has fit a verbal or statistical model to a given set of first-order data, one can then go back to the original sources of information to specify the strengths and weaknesses of one's model in more detail, with greater accuracy; even if one cannot modify one's model easily to handle the exceptions, one can try to express the information embodied in the exceptions in a more compact, more abstract form, to make life easier for those who wish to make predictions or to modify the current models in the future.

In brief, we are suggesting that descriptive research be viewed as a means to an end, with the end being prediction. A direct and total assault on the objective of prediction may indeed be a poor strategy for achieving this end. However, our success is likely to be even less, if we do not keep the basic objective fixed firmly in our minds. Every once in a while, it is important to bring together the various propositions, mathematical and verbal, which one believes to be useful in prediction, and see how effective (and consistent) they really are in coping with the overall picture. When there are major new

defects or possibilities apparent at the general level, it is important to take note of them, at that level, so that they can be used as a guide for more specialized research in different branches of political science.

Descriptive work may not only help provide the basis for evaluating dynamic models of politics; it may also help those who wish to predict politics, by telling them what the current states of the systems are to which they would like to apply the dynamic laws. The longer the policy horizon, however, the more important it is to use more general dynamic models, instead of assuming some sort of simple extension of present trends and conditions as gauged by descriptive studies. Finally, while prediction may be advocated as the primary goal of objective political science, normative political science remains another matter.

One may note, in this connection, that the attempt to maximize accuracy in <u>description</u>, by itself, leads naturally to a number of uncoordinated, specialized efforts, focused in depth on different primary sources of information(10). In <u>predicting</u> complex dynamic systems, in contrast, one finds oneself led to focus first of all on the <u>interactions</u> between the primary subsystems, at an aggregate level. Thus in order to make "interdisciplinary research" a reality in the

social sciences, it is essential that the ideal of predictive power gain at least as much credence as the ideal of descriptive finesse, in the detailed conduct of actual studies.

Fourth of all, one should note that our emphasis on prediction does not require a reduction in the rigor of thought, even if it does require that we go beyond the Aristotelian concepts of truth versus falsehood as ascertained by traditional uses of deductive logic. The mathematical theory of probability, and the Bayesian theory of inductive logic, have long provided a rigorous basis for handling models which do predict the future but which avoid the determinist's pretense of absolute certainty. Aristotelian statements, which tell us that a proposition is simply true (probability one) or false (probability zero) are simply a subset of the statements which can be expressed in rigorous probabilistic fashion. In either case, the statements that we make may well be inaccurate, if they are founded on faulty information; the language of probability, however, at least lets us express precisely how much confidence we do have in a proposition, instead of forcing us to say nothing or to exclude totally a real but less probable contingency.

When social trends will depend on long-lasting but uncertain and novel phenomena, the language of probability encourages us to escape the fallacy that definite optimistic or pessimistic predictions are somehow more informative than the truth. (See, for example, footnote (6).)

To the statistician, all these concepts have long been obvious. In verbal research, however, the classical Aristotelian procedures have remained dominant. Only in recent years has the "Bayesian school" begun to educate verbal decision-makers in the use of probability theory as a generalized language of thought(11). In section (v) of Chapter (II), we have emphasized the point that the Bayesian approach to inductive reasoning can be applied to inductive reasoning as a whole, not merely to reasoning about quantitative variables. When, in verbal research, one finds oneself dealing with the behavior of quantitative variables, such as the <u>degree</u> of popular discontent, etc., one may even go so far as to discuss the "degree of fit", the "degrees of freedom" and the "exogenous variables" of one's verbal model, on the understanding that one is expressing one's model in verbal terms only because of the lack of hard data; even then, one may want to draw together the elements of one's model, and

express them in increasingly mathematical terms, even if the parameters cannot be easily measured, in order to make its meaning more and more explicit, and in order to improve its "coherence", i.e. its completeness and its consistency.

Finally, and most importantly, our emphasis on prediction has been the driving force behind both our empirical research, and our conclusion that conventional routines for time-series analysis are inadequate. The empirical work on political science in this thesis was motivated almost entirely by the attempt to convert the Deutsch-Solow equations, mentioned in Chapter (II), into a useful tool for the prediction of national assimilation and political mobilization. We started out years ago by testing out the Hopkins routines(12), which try to estimate the coefficients of the Deutsch-Solow model from only three data points, on the assumption that the Deutsch equations are totally "true" in the Aristotelian sense; it came as no surprise to us that the resulting predictions were rather poor.

Our next step was to try out time-series multiple regression, the mainstay of "econometrics"(13), of "path analysis"(14), and of "causal analysis"(15) and

so on. It was easy enough to measure assimilation and mobilization coefficients which were significantly different from zero, and multiple correlation coefficients larger than ninety percent; these results were well within the range of what are regarded as "successful conclusions" in most quantitative political research today(16). However, our emphasis on prediction led us to look a bit more closely at these results; we wrote a new program, SERIES, to estimate the regression models, and then to test their ability to predict data across long intervals of time, intervals comparable to those tested with the Hopkins program. The errors, while less than those of the Hopkins programs, were still unacceptably large. A simple curve-fitting procedure, by contrast, was able to make predictions with less than half as much error, averaging to about 4% error over periods of time on the order of a century; this average encompasses a number of cases wherein the model was fitted to data in one period of time, and used to predict data in later periods.

Walter Isard, in his classical study of methodology in regional science(17), has made strong statements against the ability of regression models to predict the future; while his argument is phrased in

theoretical terms, the wide coverage of his studies would imply an empirical basis for his conclusions. The Brookings Institute has also reported major difficulties in the use of regression in forecasting; they found that these difficulties could be reduced by the use of an "adjustment" factor not too different in spirit from our simple curve-fitting procedure(18). Thus the empirical basis of these conclusions goes well beyond our own examples.

In our recent phase of empirical political research, we began with the hope that this weakness of regression, in estimating predictive models, could be understood within the classical and elegant framework of maximum likelihood theory, as described in section (v) of Chapter (II). Instead of questioning the classical procedures of statistics, we hoped to apply these procedures to more sophisticated models. In Chapter (III), we have noted that "white noise" in the process of measuring data can turn an ordinary "autoregressive process" into a "mixed autoregressive moving-average process." According to statistical theory, multiple regression is a good way to estimate the former process, but a bad way to study the latter.

A simple diagram can show how bad this problem might become, in practice. Given a single variable, z,

which has a true correlation of ϕ with itself across time (i.e. z(t) with z(t+1)), and a correlation of r with the measurements, x, that are made for z, we find that the correlation between x(t) and x(t+1) is due to an indirect path of correlations:

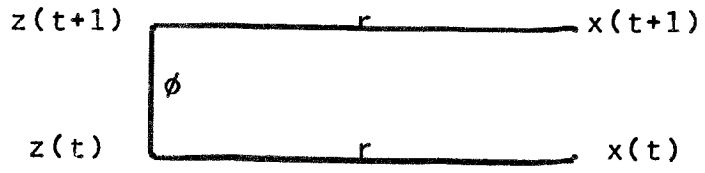

Figure V-1: Pathways of Correlation With Noisy Data

If we make the simplifying assumption (a big one) that the process is not any worse, that there is no correlation between x(t) and x(t+1) independent of this pathway, then classical theory tells us that the correlation between x(t) and x(t+1) will equal r times ϕ times r, i.e. $r^2\phi$. If regression were used to predict x(t+1) from x(t) (the observed data), the regression coefficient would equal the simple correlation coefficient, $r^2\phi$, instead of the number ϕ; yet when predictions are made over longer intervals of time, then ϕ, the coefficient of the underlying process in the real world, is the proper basis for prediction(19). This example would also appear to point to the idea that simple "path coefficients" which are not effective in prediction are not likely to represent the true

underlying relations, either.

If r - the correlation between the true variable and the measurements of the variable - were about 95%, then the observed regression coefficient ($r^2\phi$ = (.95)(.95)ϕ) would be about 10% smaller in size than the right regression coefficient (ϕ) for use in long-range prediction. Furthermore, since this 10% error would represent a general shift in the value of a coefficient, one would expect that the use of the regression model would lead to errors which <u>accumulate</u> at the rate of ten percent per time period; it is easy to see how this phenomenon alone could vitiate the predictive power of regression. In the case of a single variable, this 10% error applies to a single large correlation coefficient; therefore, one can hope that regression will at least preserve the sign of this coefficient intact. In the case of many variables, however, the 10% error would apply to a correlation <u>matrix</u>; small but critical cross-terms, on the order of $\pm$5%, might conceivably have their signs reversed, due to the spurious effects related to other, larger terms in the same matrix. After all, the importance (and much of the detectability) of such "feedback terms" lies precisely in their ability to accumulate and determine the long-term behavior of the system; it is

precisely the long-term behavior which we find poorly accounted for by regression.

In order to account for such effects, within a classical statistical framework, we have devised a new algorithm to estimate mixed autoregressive moving-average processes ("ARMA" processes) at a manageable cost. We have applied this new algorithm to the old Deutsch-Kravitz data on assimilation and mobilization in a dozen or so nations, and we have also applied it to new data on linguistic assimilation in Norway. In both cases, statistical theory indicated that the ARMA model was better than the old regression model; it indicated only a small probability, far less than 1% in almost every run, that the improvement was due to coincidence. However, when we went on to apply the test of prediction, we were quite disappointed. The ARMA model did indeed reduce prediction errors, in comparison with regression, by about 10% of the original root-mean-square average of the errors, in the case of our largest data sample; yet this is still far less than the 50% reduction achieved earlier with extrapolation.

The success of extrapolation would appear to indicate that the underlying processes are still more deterministic than either the regression or the ARMA

models were able to discover. Apparently, the measurement noise and transient fluctuations were too complicated for a "white noise" model to cope with. In retrospect, it seems clear that one should have expected precisely such a situation, in the case of both this and most other data in political science. The solution to such difficulties, in the classical philosophy of maximum likelihood, is for us to pose ever more complicated higher-order models of process noise and measurement noise. (With some other data-series, however, the ARMA model, or even the usual regression model, might be adequate.) However, the multivariate ARMA model already contains a large enough number of degrees of freedom; to double or triple the number of coefficients to estimate would put a heavy burden on all but a few very large data sets, while still compensating for only moderate complication in the noise process.

The success of simple extrapolation points to a more practical approach to prediction. It points to the possibility of "robust" estimation, of estimation techniques which can perform well despite any oversimplifications in one's original model(20); a good performance, in this context, means that the coefficients of the model are estimated in such a way

that the model will have maximum predictive power. In section (vii) of Chapter (II), we have suggested that simple extrapolation is "robust" - more robust than ARMA estimation, even apriori - because it is based on a simple "measurement noise only model," a statistical model built to extract the deterministic underlying trends (if they exist) from a process afflicted by a complex pattern of transient noise and measurement error; we have pointed out that the general dynamic feedback procedure of Chapter (II) can be used to estimate more general models of this type, economically. (It is also possible to make some allowance for process noise in an ad hoc way(21), but the best way to make such allowance while preserving robustness is unclear; there might be no general theoretical answer to this question.) We have also discussed another new technique in Chapter(II), "pattern analysis", to draw out more direct measurements of the underlying dynamic variables.

In brief: our emphasis on prediction has led us to the conclusion that statistical methods based on the concept of maximum likelihood alone are inadequate in practical empirical research. It has led us to the theoretical conclusion that predictive power itself needs to be maximized more explicitly in the model

estimation techniques available to behavioral scientists. In Chapter (II), we have suggested ways to build new systems on this principle. These systems are scheduled to be available as part of the Time-Series Processor package, designed for social scientists, in 1974 at Project Cambridge, M.I.T.

(iv) POSSIBILITIES FOR STATISTICS AS AN EMPIRICAL TOOL IN REAL-WORLD PREDICTION

Now let us come back and look more closely at the questions we started from in section (ii); let us reconsider the worries of the political actor about the use of mathematical approaches. We have dealt with these worries so far on a very basic level, on the level of defending the concept of predictive theories in the social sciences. We have emphasized the point that the explicit statistical techniques we have proposed are merely one tool among many in constructing such theories. We have implied that the choice between these techniques for constructing theories, and the verbal and Bayesian techniques, should be decided on a case-by-case and even study-by-study basis, based on

the ideal of predictive power, rather than decided by any apriori fiat in favor of one approach or the other; we have implied that statistical analyses should merge into other analyses, mathematical and nonmathematical, in such a way that political science is divided up according to interfaces between substantive issues rather than interfaces between methodological schools of thought.

All of these comments, while controversial within the domain of political science, would seem rather bland and basic to many real political actors. Most political actors would be quite willing to try any methodology that "works", at any time, without getting too committed to one methodology or another. What worries them is a question on another level: can we expect statistical methods to "work" very often, in practice?

In principle, this question can only be answered after the fact, in each case. However, there are a number of reasonable guesses one might make, based on past experience, as to the most likely areas of fruitful statistical research in the future, in political science.

First of all, we may expect to be surprised in the future, by statistical methods having a larger range of

application than one might expect apriori. Given that our present emotional expectations are built upon verbal techniques which have been thoroughly used and thoroughly developed, while our statistical techniques are only now being perfected and have barely begun to be used to construct predictive models, we may expect that the development of statistics will demonstrate a much greater value than one's intuition would indicate today.

Second, we may expect statistics to provide the basic "reality testing" for operational political theories of the quantitative type or the verbal analytic type. A simple regression analysis may make a poor test of a verbal "hypothesis", if interpreted naively. However, a full statistical analysis of a given set of variables will give a much larger quantity of information, information which the analyst should not gloss over, either in his work or in his written reports; if, in fact, it is difficult to make a connection between one's verbal theories and the aggregate, statistical behavior of the variables these theories pretend to explain, then one has much to learn in trying to explain the difficulty. Oftentimes, an "obvious" verbal theory will turn out, though true in the abstract, to require major qualification in terms

of what it tells us to expect to find in concrete data. When government policy is concerned with the concrete results themselves, these qualifications may turn out to be of central importance.

The classical theory of full-employment equilibrium, for example, depended on the idea that the supply of savings would increase with higher interest rates, just as the supply of any other commodity increases when a higher price is offered(22). Empirical studies have not refuted this idea; however, they have shown that the predictive power of interest rates in predicting savings is extremely small, while the effect of income variables, cited by Keynes, has turned out to be very large(23). Keynes himself was able to observe these effects by analytic methods alone, but major governments were very slow to change their established viewpoints despite his arguments(24); the statistical studies, by confronting people directly with the trends which had persisted up to the current day, may have been a crucial form of "reality testing" on this issue.

Third, one may hope that statistics will help illuminate the slow and stubborn trends which underly social phenomena. It has often been suggested that the most visible variables in politics - the turmoil, the yearly ups and downs in economies, alliances and wars -

are all superficial ripples riding on a deeper current. Economists have often discussed a "technological" increase in production capacity per capita, an increase which continues during recession and boom at virtually the same rate, an increase which is not fixed but which varies slowly according to uncertain causes over long periods of time. Almost any discussion of the "production function" includes reference to this "autonomous" or "technological" term(25). Sociologists like Max Weber(26), philosophers like Hegel(27) or Marx, and historians like Spengler(28), Toynbee(29) and even McNeil(30) and Eisenstadt(31) have all discussed such trends. Even in our simulation studies in Chapter (IV), we found that errors in short-term predictions were reduced far less by sophisticated analysis than were errors in medium-term and long-term prediction.

Political actors, in the short-term, are compelled to immerse themselves in details too small and too unpredictable to be dealt with effectively by statisticians. But the effectiveness of political actors also depends on their "historical vision", on their ability to judge the results of their life's work on the subsequent tide of events; long-term trends may be more deterministic, and more susceptible to mathematical analysis. In order to sort out these

underlying trends, one must somehow adjust for the existence of a great deal of short-term fluctuation. This short-term fluctuation would typically have a very complex and changeable pattern of autocorrelation; thus a complete model of the "measurement noise" process is doubly infeasible, and one must face up to the need for "robust" procedures, as described in section (iii) above and in section (vii) of Chapter (II).

After a "robust" analysis, one may find that some variables tend to follow deterministic laws, over time, but that others still involve a great deal of apparent randomness. In this case, one might expect that the political actor would have his greatest personal effect on history by trying to change the latter variables - which <u>can</u> be changed - and by aiming only indirectly at the former variables. Also, by knowing where events would be headed if he acted like the <u>average</u> political actor, an informed political actor may judge the importance of taking unusually intense actions to break out of the existing trends. Furthermore, if there should turn out to be a "crossroads" of possibilities ahead ("bifurcation", in mathematical language), such that the choice of possibilities ahead would produce very long-lived effects, while other implications of

present policy would be washed out in time by random noise, one might well choose to organize one's entire policy around the goal of moving down the right road, even if this means focusing one's attention on variables which are harder to affect.

In practice, there are serious difficulties in using statistics by themselves in dealing with this third objective. Statistics might do well in predicting the stress which will pull at the fabric of various societies; it will not do as well in predicting the ability of local political leaders to cope with the stress. Still, to know the causes and the magnitude of the stress would be interesting in any case. But there is a bigger difficulty with using statistics here. The deeper historical trends can be analyzed best if we make use of the longest possible relevant data series; yet much of our historical data base, as described by Toynbee and McNeil, involves information about civilizations which have not left us a large supply of statistical data series.

In some cases, a large supply of recent data may be enough. It may even seem superior to historical data, on grounds that it reflects exclusively modern phenomena which one would expect to continue in the future; for example, certain aspects of population

dynamics may be dealt with reasonably enough from recent data, especially insofar as the future will depend heavily on the impact of recent phenomena such as large-scale female literacy. On the other hand, if there is any truth at all in the findings of Toynbee and Spengler, then the central trends of history may include regular rises and declines(32), regular curvatures, which would be much harder to observe in short data-series - even short representative data-series - than in very long data-series. Thus the long data-series do much more than increase the number of observations and improve the accuracy of our estimates of parameters; they give us the power to deal with important qualitative effects, with whole new terms in the model, which might otherwise be missed.

Furthermore, one might expect that the future would represent a dynamic domain just as different from the present, as the present is from the past; in order to predict this domain, it may be more rational to look for regularities which have extended from the distant past to the present, and extrapolate them, rather than extrapolate models specific to the dynamic domains of the present. If, in some cases, history were dominated by large, infrequent and apparently irreversible changes, as in technology, then a longer data-base

would be even more essential, to give us an adequate sample to represent the varieties of such changes and configurations in the past; only in this way can we hope to deal with the further changes which, under this assumption, would dominate the future. (This does not, of course, entail building models tailored to shorter, well delimited periods in ancient history, as unrepresentative and restricted as recent history.) Also, to deal with the possibility that the human race might be entering a new domain of experience, totally different from any of its past history, one might simply extend the basic context of one's analysis still further, to include the more general history of species on this planet and the patterns of evolution revealed therein. The biological example of the trilobites, which became totally extinct after overspecialization, does not have a full-scale parallel in the human past; this particular example has been mentioned often in the popular press, but there may be other aspects of biological history more general and even more relevant to our own future(33). In general, it would appear futile to try to predict the fine details of a complex, natural system such as human society before we can construct first-order models which can cope with a general review of the aggregate behavior of this system

across the whole of the available data-base.

All of these visions of history emphasize that rapid periods of expansion, lasting for decades or millenia, have existed often enough in the past, and have been terminated often enough by the growth of counter trends; they emphasize that an analysis of social dynamics, based only on the data available in recent decades, may lead to a totally false picture of the possibilities which lie ahead. This difficulty certainly applies to verbal theorizing, just as much as to statistics; with verbal theorizing, however, the historical data are far more extensive for those who are willing to examine them. While we would not agree with the exact details of the theories of Spengler and Toynbee, we would consider it all the more important to describe and explain the phenomena they have discussed.

Finally, statistical methods may help on another level - as a paradigm to guide verbal research, both in general and in specific cases. We have emphasized throughout this chapter the value of verbal research, conceived as an attempt to do with verbal data what statistical research does with mathematical data. Yet even in statistics, where the methods used are spelled out explicitly in advance, we have seen that the

popular methods of analysis can be quite unrealistic and deceptive. It would seem unreasonable to expect any more, apriori, from verbal research. Indeed, we have been able to compare the two forms of research in respect to three issues: the emphasis on prediction, the willingness to deal with the longest possible time-series, and the willingness to accept noise as part of predictive theories; in all three cases, especially the first and most basic, classic verbal political science - with a handful of notable exceptions - does not appear to have grounds to claim superiority in its attitudes, at least not in the parts of the literature with which we are familiar(34).

In this situation, the methodological advances in statistics - which are well-defined, and which can be consolidated - can be of major value in educating the verbally-oriented political scientist. The concept of stochastic predictive theories may seem reasonable in the abstract to the verbal political scientist; however, when he tries to translate this idea into a strategy for his own research, it would not be surprising if the difficulty of doing so brought him to withdraw back to Aristotelian procedures. Indeed, the classic attempt to track down the long-term "causes" of historical events is, as we have mentioned above, very

closely linked to the search for dynamic models. However, the concept of "cause" is a weak enough paradigm that historians have often found themselves forced to admit "multicausality"(35), and then to withdraw into more descriptive, more "objective" questions. Furthermore, historians have often admitted themselves to be intrigued by the "great 'what if' questions of history", such as, "What would have happened if the Spanish Armada had won in 1588?"; yet such questions - which clearly call for the use of some kind of predictive model - have been dismissed as speculative(36).

In short, there would appear to be a need for a more durable paradigm in analytic verbal research. If verbal social scientists can become more and more familiar, on an intuitive level, with the concrete methods of statistics, in coming to grips with concrete data, then they may be able to develop a clearer and clearer picture of what it means to search for robust predictive models, mathematical or verbal. Also, they may be expected to learn to appreciate the value of treating quantitative variables as such, even in verbal discussion, rather than reducing them to such possibilities as "high" and "low"(37).

The value of statistical methods as a paradigm for

verbal research may be especially great in those cases where statistics can deal explicitly with some, but not all, of the historical data of interest. One can imagine two ways in which this value would be felt, as statistics are brought to bear on those data which are available.

First of all, those people doing the statistical research would have to choose a set of variables to use, in developing predictive models. When predicting a system like a missile, made up of five major subsystems or so, one's primary concern in making medium-term predictions is with the "overall system," with the system made up of the interactions between the five major subsystems. Similarly, when asking for long-term prediction of a statistical system, made up of five clusters of heavily intercorrelated systems of variables, one would normally start out by aggregating each of the clusters, by use of factor analysis, pattern analysis or other procedures, and then studying the relations between the aggregate variables. To try to predict where a missile will go, by predicting what each of the subsystems would do in total isolation from each other, is to ignore the most important functional relations. Statistical analysis, by drawing us away in concrete cases from a fixation on the internal dynamics

of specialized subsystems, may help bring us back to studying the broad structure of interfaces and multiple subsystems which is crucial to even a first-order prediction of human societies. Then, when researchers go back to look inside the subsystems, we may hope that they will focus more attention on those questions about the interfaces which appeared important at the global level. The "unity of science" may be a debatable proposition when applied to predictive models of, say, biological systems and astronomical systems. However, when one is trying to predict a single, highly integrated system, the need for interdisciplinary unity becomes overwhelming. When small feedback terms from one subsystem to another can have overwhelming effects in determining system behavior, it is essential to try to measure the aggregate behavior directly.

In the next phase, <u>after</u> the statistics have pointed to concrete interdisciplinary effects, human verbal knowledge can go on to explain and to qualify these conclusions. One might, on verbal grounds, regard the conclusions as misleading, as oversimplified, or as one-sided in their emphasis; in any case, however, even to discuss these conclusions intelligently, one must try to discuss, on the basis of <u>verbal</u> knowledge, why one would expect certain

correlations and certain dynamic patterns to work out as they do. One would have to discuss the "true" dynamic relations between different subsystems, in order to express what one feels are wrong with the statistics. One would be tempted to engage in "hypothetical" or "analytic" modelling - to suggest what would have happened to the statistics if one had included, as hard data, certain variables for which hard mathematical data do not happen to exist. One would focus one's attention on the variables of central interest, rather than lose oneself in a morass of unrelated higher-order vicissitudes. In brief, one might acquire the momentum necessary to launch into a full-scale verbal dynamic analysis, without crashing back under the weight of pure classical traditions.

Before we close this discussion of the value of statistics to political analysis, it may be worth noting that important applications may also exist in politics proper. Early in Chapter (II), we mentioned the possibility of a growth in ecological and sociological models, based on the vast accumulation of data by earth satellites. Many of the difficulties cited above - particularly the dominance of verbal data over quantitative data - would not apply in this case; also, our emphasis on feedback effects and

interdisciplinary research would apply more than ever. Given the importance of world balances of agricultural production and population, given the dangers of ecological catastrophe through high stress and imbalance in such systems, and given the possibilities here for treaties to set up a coordinated global system to monitor and help control these systems from space, the practical political scientist would have good reason to think about these applications. This is doubly true, insofar as the development of these applications may be far from automatic.

(v) BEYOND NAIVE EMPIRICISM: ADAPTING OUR IDEAS TO FILL THE GAP LEFT BY STATISTICS

Now let us return once more to our starting point, to the worries of the political actor about using mathematics. (See section (ii).) We have dealt with these worries on two levels: (i) the level of defending the notion of prediction; (ii) the level of describing the practical applications of statistical modelling to political prediction. We have emphasized the empirical approach, in both verbal and statistical

research.

On another level, however, the political actor might question the idea that empirical approaches are enough, when the subjects of one's investigation are intelligent human beings. In particular, the political actor would have to reconcile the use of objective methods to predict other actors, while preserving the sense of free will in making his own decisions. On a primitive level, this paradox poses no difficulties at all to the political actor; it is easy to conjure up the image of a fast-dealing political hack, working for a city machine, gleefully pushing people around as if they were buttons on a pin-ball machine. At a more advanced level, politicians find that they can predict people better and influence their actions more constructively by exploiting empathy, by using their own reaction patterns as a kind of analogue model to predict the reactions of others. Thus there are the old, persistent adages:"If you want to predict what a man will do, try to put yourself in his shoes," and, "If you were a ... what would <u>you</u> do?" This procedure is particularly effective when the political actors come from the same background as the people they are predicting, when their background is cosmopolitan, or when their reaction patterns are defined at a general

enough level to make it easier to imagine how another person would really respond to a situation very different from that of the political actor himself. (Empathy may also be used, of course, as a tool in thinking of approaches one might borrow from others in coping with one's own problems.) When a political actor oscillates between thinking about others "subjectively", in terms of empathy, and thinking about them "objectively", in terms of predictive empiricism, a conflict emerges, long before the use of statistics as such arises. Rationality, and the acknowledgement of others' capacity for rationality, would appear to allow no escape from this conflict; as long as we have two distinct sources of information from which to predict the behavior of people, we must live as best we can with the conflicting predictions, <u>while trying to reconcile them by concrete improvements in the concepts we use on both sides</u>.

Predictive statistical models, like empirical verbal models, cannot directly express the insights derived from "empathy"; in particular, they cannot express the insights derived from acknowledging the intelligence of other human beings(38). This limitation may be of enormous importance to the practical political actor. On the other hand, the related

mathematical concept of maximizing a cardinal utility function expresses the idea of human intelligence, more vividly and more precisely than the usual verbal formulations. The arguments of Von Neumann(39) and of Raiffa(40) in favor of this concept require little more than logical consistency on the whole, in the ultimate values that the individual pursues; while the serious political actor would normally admit that he sometimes acts stupid, and sometimes acts at cross-purposes against himself, especially when limitations on time and on knowledge constrain his detailed decision-making, he would rarely consider such mistakes as a matter of fixed or deliberate policy. Often, when the political analyst would accuse him of indulging in irrationality, he would have a counterargument of his own, based on the knowledge and concepts available to him at the time of his decision.

If we agree with Raiffa, then, that the maximization of cardinal utility is "valid" as a foundation for most political decision-making, we find ourselves led to important conclusions about political analysis too. First of all, we find ourselves re-emphasizing the point that verbal research may be regarded as an attempt to perform valid statistical inference, accounting for data which is less structured

and less manageable than the usual statistical time-series. Within Raiffa's framework, the basic questions one asks are quantitative in nature; e.g. - "If we carry out action A, how much will it cost, how much do we gain, and what are the probabilities that we will succeed?" More generally, Raiffa would have us ask,"If we carry out action A, starting from situation B, what is the distribution of probabilities attached to the different possible levels of cost and to different possible outcomes? How much do we expect to gain from each of the possible outcomes, if our subsequent strategy is optimal?"

In each case, we do the best we can to estimate these quantities on the basis of the available verbal information; thus the research carried out on that information, is carried out for the purpose of extracting the most accurate possible statistical information. We also account for the intelligence of other actors. We also account for more direct quantitative evidence, whenever we can find it. From most sources of information, we expect to get probabilistic indications of various kinds, never certainties. Through practice, we may hope to learn more and more the art of formulating accurately the interrelated patterns of statistical implications of

our verbal knowledge, and to reduce the losses in translation which always intervene in going from raw observation to decision.

Second - and more important - the concept of utility maximization offers us an idealized model of intelligent decision-making, for use in the prediction of other political actors. At first glance, this concept may sound rather culture-bound. However, the concept of utility function is very generalized in the range of concrete behavior it can include. One can imagine all sorts of different utility functions. One can imagine many different levels of knowledge and aptitude brought to bear in maximizing utility functions. One can even imagine different levels of basic cognitive structure, as suggested by Piaget(41) and by ego psychiatrists(42), levels which one may hope either to remember or to advance to. One can imagine states of short-term psychological disequilibrium, where a political actor does not yet take the actions best suited to maximizing his utility function, because, on some level, he has not yet become aware of the possibilities. (The detection of such disequilibria is particularly important to political actors whose job is to persuade others to change their course of action.) Thus, <u>starting</u> from the concept of utility

maximization, one can approach empirical reality bit by bit, by using both empathy and empirical data to add qualifications to one's view of other actors as ideal decision-makers. Even as one adds qualifications, however, one can continue to insist that all interpersonal differences in personality be analyzed in terms of the current state parameters of a system which obeys the same general dynamic laws as one's own mind, and which is capable of changing its state parameters as a result of the general learning capability shared by all humans(43).

From a theoretical point of view, the choice of starting point is not a matter of mere bookkeepping; it defines one's implicit "prior probability" distribution, as described in section (v) of Chapter (II). From a practical point of view, this procedure can help us avoid rigid stereotypes of other political actors; it can help us remember that they, too, have a capacity for change, and that the likely directions of change are not entirely random. In any case, this procedure allows us to make use of both major sources of information, information derived from empathy and information derived from more objective data. This procedure also suggests that the procedures mentioned in section (x) of Chapter (II) for utility maximization

might be used, not just as a technique for analyzing decision-making systems, but as the basis for substantive models of such systems.

The use of utility maximization as a model of decision-making has already led to a number of practical applications, notably in game theory and in microeconomics. The concept of ideal utility maximization predicts the behavior of an actor conditional upon the information that he has available to him; thus it may lead to an implicit model, a model defined in terms of variables which are not directly observable to other actors. In terms of behaviorist attitudes, this is a major liability, insofar as it makes it much more difficult to predict behavior concretely; on the other hand, such implicit models may allow us to infer something about the hidden variables from the overt behavior.

In the case of microeconomics, one does not attempt to predict the actual levels of steel production, etc., at least not in the early stages of research; instead, one defends the proposition that the levels of steel production will be equal to whatever level is necessary in order to maximize some kind of utility function, if the decision is made by an economy which enjoys perfect competition(44). (Strictly

speaking, however, economists now tend to avoid the concept of "social utility function", on grounds that they can deduce similar conclusions from weaker versions of the same criterion.) One might well have attacked this theory, in its early stages, as an unscientific - though mathematical - exercise in propaganda. However, as the theory was developed, it turned out to be a powerful framework for evaluating the inefficiencies produced by situations of imperfect competition in the real world(45); it has been used to analyze the effects of taxes and labor laws(46) on economic efficiency; it has led to the development of Lieberman's principles of economic organization(47), now a mainstay of the Soviet economy(48). Microeconomics, initially an isolated and essentially unempirical theory, has turned into a powerful mathematical framework for analysis, a framework allowing the useful bringing together of vast quantities of empirical data, a framework important to both the prediction and the comprehension of economic phenomena.

Yet all of this success was based on a static concept of utility maximization, a concept of optimal equilibrium, related to the classic concepts of Lagrange(49). Since then, Norbert Wiener(50) has

discussed the more modern, more powerful dynamic theories of maximization, which he would consider a substudy of "cybernetics." He has suggested that this body of theory be applied to the human brain(51). Karl Deutsch(52) has gone on to suggest that cybernetics may also be applied to political science. Considering the power that a primitive, static concept of maximization has had, over decades, in economics, these suggestions would appear to make a great deal of sense. Unfortunately, these ideas are caught between the "mighty opposites" of modern methodology - the behaviorists, who would demand quantitative empirical proof that the initial model predicts all the variables in detail, and the traditionalists, who would not have patience with the mathematics. Also, in the last few years, the relevant phase of "cybernetics" has been renamed "control theory." We have discussed the value of "control theory" (i.e. of optimization techniques) in Chapter (II) as a tool in analyzing social systems; however, control theory may also be used itself as a normative model, of the processes which allow human societies - or even the human brain itself - to function. (See note (53) for more concrete possibilities.) As with microeconomics, one will expect to find that the real systems involve imperfections and

approximations to the optimum.(54). However, one may also find it interesting to be able to see where these imperfections are, and to appreciate the capacities that human beings and political societies - like economies - do have, to cope with data on a scale far beyond the capacity of present-day computers. Insofar as one agrees with the traditionalist that the human being is still the most relevant unit of analysis in politics, one can try, in the future, to expand the interface between cybernetics in psychology and cybernetics in political science.

In summary, we have concluded that the mathematical methods outlined in Chapter (II) can indeed be applied to political science, but that they should always be considered as only one branch of a more complex, integrated system of analysis, oriented towards the goal of prediction; the Bayesian philosophy of utility maximization and conditional probabilities could play a central role in organizing this system of analysis, but the behaviorist philosophy of total empiricism does not have the power to account for major parts of this system, parts which appear essential in the last part of this chapter.

FOOTNOTES TO CHAPTER (V)

(1) One might take up considerable space discussing this particular point. Traditional Lanchester's Laws for ancient warfare indicate an equal number of deaths on both sides, regardless of concentration; thus, they tend to imply no possibility of strategy in such warfare. On the other hand, the "Laws" usually have the disclaimer "ceteris paribus" attached, implying equal levels of material and social "technology." If "social technology" includes superior strategic ability on the part of a commander, like Caesar, then the laws become a poor guide for the would-be superior strategist. We have limited our statement here to the claim that Caesar won his victories by <u>evading</u> Lanchester's Laws, by avoiding the necessity for attrition or perhaps by exploiting the loopholes in the laws, rather than invalidating them; this much, at least, seems fairly clear to us from Caesar's account itself.

Liddell-Hart, in his classic military textbook, <u>Strategy</u> (Praeger, NY, Second Edition, 1967), cites (p.338) Caesar's Ilerda campaign, Cromwell's Preston campaign, and a few others, as the classic bloodless victories; he goes on to write, on p.339:"While such bloodless victories have been exceptions, their rarity enhances rather than detracts from their value - as an indicator of latent possibilities, in strategy, and grand strategy. Despite many centuries' experience of war, we have hardly begun to explore the field of psychological warfare. From a deep study of war, Clausewitz was led to the conclusion that - 'All military action is permeated by intelligent forces and their effects.'"

(2) Liddell-Hart generally prefers to talk about the "indirect approach" and the "unexpected" more than the use of imagination, but clearly the former require the latter. Hart writes (ibid) on p.342:"A more profound appreciation of how the psychological permeates and dominates the physical sphere has an indirect value. For it <u>warns us of the fallacy and shallowness of attempting to analyze and theorize about strategy in terms of mathematics</u>. To treat it quantitatively, as if the issue turned only on a superior concentration of forces at a selected place, is as faulty as to treat it geometrically..." Also, on p.162, in the

section "Conclusions from Twenty-Five Centuries", Liddell-Hart discusses the characteristics of the more usual, bloody victories:"... scanning, in turn, the decisive battles of history, we find that in almost all the victor had his opponent at a psychological disadvantage before the clash took place... most of the examples fall into one of two categories... described in the words 'lure' and 'trap'." See also note 1. The full weight of these objections will not be dealt with until section (v) of our text.

(3) In writing this, and remembering some of the interesting generalizations we had heard from pure verbal political science, it was difficult to overcome the selective memory and face up to the overall methodological views still prominent in the field. But a quick review soon set our memory straight. For example: "It should be noted that the emphasis here is on deduction, not on induction. In the words of another participant in the seminar, Professor S.E. Finer, we are making an attempt at 'describing the political possibilities.' Considerable emphasis should be put on the word '<u>describing</u>': we remain in the humble sphere of description and do not attempt to rise to the more lofty one of speculation." (p.40 of "General Methodological Problems", by Gunnar Heckscher, in <u>Comparative Politics</u>, Eckstein, Harry and Apter, David E., eds., Free Press of Glencoe, 1963.) In historical research, the problem is more serious, as: "My principles and methods of research and writing were largely worked out unconsciously, through listening to excellent teachers and following the best models... The historian has both the right and the duty to make moral judgements. He should not attempt to prophesy, but he may offer cautions and issue warnings." (p.44-45, <u>Vistas of History</u>, Samuel Eliot Morison, Knopf, NY, 1964.) One may ask what the warnings are supposed to be based on, if not on probabilities of undesirable events conditional upon certain policy decisions; also one may question whether methodological decisions ought to be based on unconscious factors. See also notes 10 and 36.

(4) Spengler, Oswald, <u>The Decline of the West</u>, Knopf, NY, 1926, translated from the 1918 original by Charles Atkinson, p.106-107: "The aim once

attained - the idea, the entire content of inner possibilities, fulfilled and made externally actual - the Culture suddenly hardens, it mortifies, its blood congeals... This - the inward and outward fulfillment, the finality, that awaits every living Culture - is the purport of all the historic "declines", amongst them that decline of the Classical which we know so well and fully, and another decline, entirely comparable to it in course and direction, which will occupy the first centuries of the coming millenium but is heralded already and sensible in and around us today - the decline of the West." p.109-110: "Every Culture, every adolescence and maturing and decay of a Culture, every one of its intrinsically necessary stages and periods, has a definite duration, always the same, always recurring with the emphasis of a symbol."

(5) Toynbee, Arnold J., A Study of History, abridgement of Volumes I-VI, Oxford U. Press, NY, First American Edition, Fourth Printing, 1947, p.244:"The problem of the breakdowns of civilizations is more obvious than the problem of their growths. Indeed it is almost as obvious as the problem of their geneses. The geneses of civilizations call for explanation in view of the mere fact that this species has come into existence and that we are able to enumerate twenty-six representatives of it - including in that number the five arrested civilizations and ignoring the abortive civilizations. We may go on to observe that, of these twenty-six, no less than sixteen are now dead and buried." p.245:"If we accept this phenomenon as a universal token of decline, we shall conclude that all the six nonWestern civilizations alive today had broken down internally before they were broken in upon by the impact of Western civilization from outside... For our present purposes it is enough to observe that of the living civilizations every one has already broken down and is in process of disintegration except our own." p.253-254:"The metaphor of the wheel in itself offers an illustration of recurrence being concurrent with progress... Thus the detection of periodic repetitive movements does not imply that the process itself is of the same cyclic order as they are. On the contrary, if any inference can historically be drawn from the periodicity of

these minor movements (such as the rise and decline of Graeco-Roman civilization), we may rather infer that the major movement which they bear in mind is not recurrent but progressive." (Comments in parentheses inserted by us.) Also, in various places, Toynbee emphasizes both scholastic rigidity and corruption as symptoms of decaying civilizations; he hints at a different, less charitable explanation of the common methodological difficulties we have mentioned in the text. However, even if Toynbee's explanation has some truth in it, a reduction in the cost of adhering to good methodology should still facilitate more worthwhile research.

(6) Turner's theory is well-known as an attempt to articulate the factors which caused American progress in the last few centuries; however, it is not only interesting in its own right, but is an example how such ideas can be useful sometimes to those trying to decipher more general laws of history, which, in turn, may be useful to present policy-makers. Walter Prescott Webb, in "The Frontier and the 400-Year Boom", p.136, writes in comment on Turner's ideas:"Assuming that the frontier closed about 1890, it may be said that the boom (in all of Western civilization) lasted approximately four hundred years. It lasted so long that it came to be considered the normal state, a fallacious assumption for any boom. It is conceivable that this boom has given the peculiar character to modern history, to what we call Western civilization." (Article by Webb located in Taylor, George R., ed., The Turner Thesis: Concerning the Role of the Frontier in American History, Heath Co., Lexington, Mass., Third Edition, 1972. Webb goes on to suggest that the search for "new frontiers" is essentially an irrational, desperate attempt to preserve a dying enterprise; however, his assumption that new foci of economic development cannot be found does not allow for some of the possibilities of technological progress over the next few decades. Over centuries, the limits of the earth itself may be expected to prevent unlimited growth; on the other hand, when one speaks in terms of centuries, one cannot entirely rule out the possibility of developing economic activities beyond the planet earth itself. Certain aspects of economic and technological growth depend on large numbers of

independent random disturbances, which, on the whole, may accumulate and be subject to accurate prediction by statistical procedures or the verbal equivalent; however, the historic development of nuclear power, for example, or the future possibilities of elementary particle physics and nonequilibrium (nonlocal) thermodynamics (see note 51), involve a more sweeping form of prior ignorance, which translates into probabilities far from one or zero and whose values may change according to government or even individual decisions. At any rate, it is possible that the ideas mentioned by Webb may have application to other parts of the historical data base, beyond the West.

(7) Hegel, "The Philosophy of History", excerpted in The Philosophy of Hegel, Carl Friedrich, ed., Modern Library, NY, 1954, p.21-22: "The Principle of development contains further the notion that an inner destiny or determination, some kind of presupposition, is at the base of it and is brought into existence. This final determination is essential. The spirit which has world history as its stage, its property and its field of actualization is not such as would move carelessly about in a game of external accidents, but is instead the absolute determining factor." p.23:"World history presents therefore the stages in the development of the principle whose memory is the consciousness of freedom..." Stages then listed.

(8) Schwinger, Julian, Particles, Sources and Fields, Addison-Wesley, Reading, Mass., 1970, Preface, especially paragraph two of preface. The most phenomenological approach in basic physics, the "S-matrix" approach, restricts its attention to describing the S-matrix. This matrix is defined as the matrix which predicts all scattering results, for all possible scattering experiments in high-energy physics;these, in turn, constitute the vast majority of high-energy data.

(9) In quantitative political science, the obvious reference here is to Blalock, Hubert M. Jr., Causal Inference in Nonexperimental Research, U. of North Carolina Press, Chapel Hill, 1964. Hayward Alker also recommends: Simon, Herbert, Models of Man, Wiley, NY, 1957, Part I, and Dahl,

R.A., "Cause and Effect in the Study of Politics and Discussion", in Cause and Effect, D. Lerner ed., Free Press, New York, 1965.

(10) H.R. Trevor-Roper, the noted traditional historian (and antagonist of Toynbee), writes, on p.vi. of Historical Essays, Harper and Row, NY, 1966: "It is perhaps anachronistic to write of a historian's philosophy. Today most professional historians 'specialize'. They choose a period, sometimes a very brief period, and within that period they strive, in desperate competition with ever-expanding evidence, to know all the facts. Thus armed, they can comfortably shoot down any amateurs who blunder or rivals who stray into their heavily fortified field; and, of course, knowing the strength of modern defensive weapons, they themselves keep prudently within their own frontier. Theirs is a static world, a Maginot Line, and large reserves which they seldom use; but they have no philosophy. For a historical philosophy is incompatible with such narrow frontiers." Note 36 and note 3 are also interesting in this connection; also, Pages V-34 and V-35 of our text consider interdisciplinary effects in somewhat more detail.

(11) Raiffa, Howard, Decision Analysis: Introductory Lectures on Making Choices Under Uncertainty, Addison-Wesley, Reading, Mass., 1968. This book, at least, is clearly intended to communicate to a broader community. The philosophical foundation of this view of probability is described in Kyburg, Henry E. Jr., and Smokler, H.E., eds., Studies in Subjective Probability Wiley, NY, 1964. Anatol Rapoport has criticized the abuse of probabilistic concepts by decision-makers who do not fully understand them; see his Strategy and Conscience, Harper and Row, New York, 1964, especially Chapter 10. Nevertheless, a false estimate of uncertainty may be less dangerous than a forced choice of absolute certainties; Raiffa, in a memo co-authored with Marc Alpert, has discussed in detail the problem of educating and "calibrating" decision-makers (i.e. compensating for their overconfidence), to estimate probabilities more realistically. (See "A Progress Report on the Training of Probability Assessors," available in 1971 as an unpublished manuscript from the office of Prof. Raiffa in the Littauer Building, Harvard

U.) Still, Rapoport's comments on the hazards of mathematical approaches are well worth noting, for those who would want to use such approaches.

(12) The programs EVAL, DIFF and DELTA of Raymond Hopkins were made available to us by Prof. Deutsch, Dept. of Government, Harvard. They have been described in Hopkins, Raymond, "Projections of Population Change by Mobilization and Assimilation," Behavioral Science, 1972, p.254.

(13) Johnston, J., Econometric Methods, McGraw-Hill, NY, 1972, Second Edition, p.ix: "The purpose of this book is to provide a fairly self-contained development and exploration of econometric methods... It is divided into two parts. Part 1 contains a full exposition of the normal regression model. This serves as an essential basis for the theory of econometrics in Part 2."

(14) It was surprising to find the phrase "path analysis" so rare in books up to 1973. In 1971, we discussed the subject at length with Prof. Alker, at MIT, one of the main exponents of this approach, with Prof. Raymond Tanter then of the Center for Research in Conflict Resolution at the University of Michigan, and with students taking "path analysis" as a subject in the Inter-University Consortium for Political Research. In all cases, it was clear that "path analysis" was intended as a kind of refined "causal analysis", using regression coefficients (or time-series regression coefficients, in the sophisticated versions?) as indices of the size and direction of influence. Simon, Blalock and Boudon have also been associated with "path analysis," in discussions at various universities.

(15) See Blalock, note 9. Simple correlation was usually used in "causal analysis" based on Blalock; this is the univariate special case of regression analysis.

(16) As an arbitrary example, picked from a good anthology of papers in this field, consider Singer, J. David, ed., Quantitative International Politics, Insights and Evidence, Free Press, NY, 1968, tables of results on p.278-281, p.112, p.152-153, p.199, p.205, p.65, p.232. Statistical significance scores ("p") here often run to ".05",

or even to as poor as ".10".

(17) Isard, Walter, Methods of Regional Analysis: An Introduction to Regional Science, MIT Press, Cambridge, Mass., 1963, Third Printing, p.22.

(18) Duesenberry, J.S., Fromm, G., Klein, L.R., and Kuh, E.H., eds., The Brookings Model: Some Further Results, Rand-McNally, Chicago, 1969, p.296-297. The full regression model, justified by solid significance indications, did not perform as well as a "condensed" - dramatically reduced - model, at first. Then "adjustments" were applied, which dramatically improved the fit; these adjustments seemed to entail multiplying each coefficient by a constant, suitable to adjust the predictions of each variable to the right level in the first-half test data. On p.298 they caution, even still,: "If the model does indeed suffer from omission of important but slowly-changing variables, then it is probably not very useful for long-run analysis or projection." Estimating or adjusting coefficients to maximize predictive power directly, over the trial data, is the essence of our proposal in sections (vii) and (xi) of Chapter (II) of this thesis.

(19) From Chapter (III), "ϕ" is simply the matrix "θ" in the univariate case; given past error levels, a(t-k), it is the best basis even for predicting x(t+1). However, looking at x(t+n) from x(t) only, we get a long path of correlations multiplying out to $r\phi^n$... thus to get the optimal prediction of x(t+n), one multiplies one's prediction of x(t+n-1) by ϕ.

(20) See note 19 of Chapter (II) of this thesis.

(21) See section (xi) of Chapter (II) of this thesis, for a way to introduce a kind of "interest rate" or "discount factor", to predictions of more distant times in the future. Such procedures may be unavoidable when a small amount of process noise does exist, and does accumulate through time.

(22) McCracken, Harlan L., Keynesian Economics In the Stream of Economic Thought, Louisiana State University Press, 1961, p.51: "Perhaps one of the finest contributions Keynes made to economic

theory and economic policy has been on the subject of investment. According to previous classical analysis, savings, investment, and the rate of interest all fitted into the standard pattern of demand, supply and price. A high rate of interest increased savings and decreased investment, while a low rate of interest decreased savings and increased investment, so it was a function of price - the rate of interest - to gravitate to the equilibrium point where saving equalled investment. There would be no such thing as over-saving or underinvestment (i.e. depression), as they were continuously being brought into balance by an automatic regulator.

For Keynes classical interest theory was in error at two basic points. First, while a priori reasoning leads to the natural conclusion that a high rate of interest stimulates saving and a low rate reduces saving, a posteriori evidence..." Comments in parentheses our own.

Gardner Ackley describes equivalent ideas in Macroeconomic Theory, McMillan, NY, 1961, (Twelfth Printing 1967) p.154-155: "Wicksell's analysis (the classical analysis)... gave us, as has been stressed, a rudimentary theory of the aggregate demand for goods. This demand consists of two main divisions: consumer demand and investment demand. Each of these demands was conceived to be interest-elastic: the lower the interest-rate, the greater the investment demand; and the greater the consumer demand, too (the latter idea is, of course, merely a restatement of the idea that saving depends negatively on the interest rate)... If either type of demand declined, the resulting fall in the rate of interest would stimulate them both, and shift resources to the one which had not declined. If, however, for any reason (particularly expansion or contraction of the money supply by the banks) the rate of interest were prevented from performing this regulatory function, aggregate demand... would be altered... But if wages and prices should not decline (enough)... Workers would become unemployed, and real as well as money income would be cut."

(23) The history of these studies is rather complex. The major initial study, by Simon Kuznets, National Income: A Survey of Findings, NBER, NY, 1946, uses technical language difficult to summarize here. Elizabeth W. Gilbey, the Economics

of Consumption, Random House, NY, 1968, p.25:"In attempting to test this hypothesis, contradictory results arose from the use of time-series and cross-sectional data. Simon Kuznet's study of data going all the way back to 1870 showed that the percentage of aggregate income saved had in fact remained constant in the United States." (The contradiction involved the distribution of saving across households.) A more recent survey is Patinkin, Money, Interest and Prices, Harper and Row, NY, Second Edition, 1965, p.651-664. Patinkin discusses largely the "wealth effect", based on the "Pigou effect", a more recent attempt to resurrect classical ideas; on p.656 and 657 Patinkin cites numerous studies which measure wealth as real assets times interest rates. On p. 663 he describes his own results for "beta" and "alpha", the former which he equates with "YL"(income), and the latter, at the top of p. 659, defined as beta times interest rate. Thus the latter results explicitly measure the hypothesis that interest rates affect the percentage of income saved, while the former do measure something closely related; also, the studies of Goldsmith cited by Patinkin reaffirmed the idea of a "constant saving-income ratio." Patinkin describes his own studies, and some of the previous studies, as showing large and significant effects by variables derived from interest rates; however, the actual regression coefficients of alpha ran to .04-.08, at the most, much smaller than the coefficients of beta, which was already a larger number to begin with.

(24) Many would identify the "liberal" Roosevelt with the "liberal" Keynes. However, US GNP data indicate rather strongly that the main recovery from the Depression coincided with major military spending induced, not by economic theory (though Keynes' theory might have recommended it, given no alternative spending options on the same scale), but by World War II. Keynesian theory, in many respects, was not fully accepted in the US until John Kennedy became president. Schlesinger, Arthur F., A Thousand Days, Houghton-Mifflin, Boston, 1965, p. 1005: "The (taxcut) bill made slow progress through Congress. Public reaction at first was muted. Kennedy used to inquire of the professors of the Council what had happened to the several million college students who had

presumably been taught the new economics...Still... on September 25, 1963, the worst was over... The Yale speech had not been in vain; and the American government, a generation after General Theory, had accepted the Keynesian revolution." Regarding Keynes and Roosevelt, Schlesinger has written, in The Age of Roosevelt, Houghton-Mifflin, Boston, 1966, Vol.III, p. 236: "The First New Deal, in the main, distrusted spending. Its conservatives, like Johnson and Moley, were orthodox in their fiscal views and wanted a balanced budget; and its liberals, like Tugwell and La Follette, disliked spending as a drug which gave the patient a false sense of well-being before surgery could be completed." p.403: "Shortly after Roosevelt's inauguration (1933) Keynes spoke once again in a brilliant pamphlet called 'The Means to Prosperity.' Here he argued with new force and detail for public spending as the way out of the depression. Employing the concept of the 'multiplier', introduced by his student..." p.404: "'Unfortunately,' Keynes wrote in April 1933, 'it seems impossible in the world of today to find anything between a government which does nothing at all and one which goes right off the deep end!'" p.405: "... on May 28, 1934, Keynes came to tea at the White House. The meeting does not seem to have been a success." p.406: "...to Frances Perkins Roosevelt complained strangely, 'He left a whole rigamarole of figures. He must be a mathematician rather than a political economist.'"

(25) Solow, Robert M., "Technical Change and the Aggregate Production Function", Review of Economic Statistics, 1957, p.312-320. Solow writes:"Not only is delta A over A (the percentage increase in the autonomous term) uncorrelated with K/L, but one might almost conclude from the graph that delta A over A is essentially a constant in time, exhibiting more or less random fluctuations about a fixed mean." Looking at figure 3, one notes a possible exception to this, which Solow admits is a very tentative conclusion: the growth of this term might have actually been faster, slightly, during the depths of the depression (actually, lagging it by three years in the graph), than under normal conditions.

(26) Essays in Sociology, from Max Weber, introduced by Talcott Parsons. Weber's concept of "rationalization", as a trend extending from the "'Concept' of Plato" to the "cage" of modern machine civilization, does amount to a long-term vision of history.

(27) See note 7.

(28) See note 4.

(29) See note 5.

(30) McNeil, W.H., The Rise of the West, Mentor, NY, 1965. Aside from the title, the themes are a bit too complex to summarize here. Many of them are reminiscent of Turner, note 6. Throughout the book, however, McNeil does keep returning to the theme of human societies adapted to the pastoral niche as providing the soil on which new civilizations may develop or to which old civilizations may spread, as the old heartland decays.

(31) Eisenstadt, S.N., The Political Systems of Empires, Free Press of Glencoe, 1963. Chapter 2 attempts to explain the "universal states" of the Toynbee and Spengler theories, almost the same societies.

(32) See notes 4 and 5.

(33) Lorenz, Konrad, On Aggression, Harcourt Brace and World, NY, 1966, translated by M. Wilson. This source is already popular among some political scientists. Just as relevant may be Simpson, George Gaylord, The Major Features of Evolution, Columbia University Press, NY, 1953, p.391: "The populations making a quantum shift (e.g. evolution of human intelligence) do not lose adaptation altogether; to do so is to become extinct. It is also clear that the direction of change is adaptive, unless at the very beginning... Yet the very fact that selection pressure is strong can only be a concomittant of movement from a more poorly to a better adapted status. Selection is not linear but centripetal when adaptation is perfected. It is the 'stabilizing selection'... The quantum change is a break-through from one portion of stabilizing selection to another." The

Indian caste system, or the early Caribbean system of Carib predators and Arawaks, are interesting examples of centripetal development among humans in relatively static ecologies/economies. p.392: "Quantum evolution usually is and at some level it may always be involved in the opening or so-called 'explosive' phase of adaptive radiation. The relative rapidity with which a variety of adaptation zones are then occupied seems quite inexplicable except by a series of (?) and also, often, successive quantum shifts into the varied zones. The rates thereafter slow down."

(34) See notes 37, 3 and 10. Also: Aron, Raymond, Main Currents in Sociological Thought, Vol. I, Basic Books, 1965, translated by Harold Weaver, p.4: "American sociologists, in my own experience, never talk about laws of history, first of all because they are not acquainted with them, and next because they do not believe in their existence."

(35) The "multicausal approach" has appeared in historical research, if our memory is correct, but the sociologists - who find it harder to retreat into simple narrative - have spoken much more about the idea. See, for example, Vernon, Glenn M., Human Interaction: An Introduction to Sociology, Ronald Press, NY, 1965, p.30 and p.80-81 especially. See also MacIver, Robert M., Social Causation, Ginn and Co., Boston, Mass., 1942.

(36) Trevelyan, G.M., An Autobiography and Other Essays, Longman, Green and Co., 1949, London, p.91: "The endlessly attractive game of speculating on the might-have-beens of history can never take us very far with sense or safety. For if one thing had been different, everything would thenceforth have been different - and in what way we cannot tell... As serious students of history, all we can do is to watch and to investigate how in fact one thing led to another in the course actually taken. This pursuit is rendered all the more fascinating and romantic because we know how very nearly it was all completely different. Except perhaps in terms of philosophy, no event was 'inevitable.'" Historians have often discussed the "turning points," times when the susequent course of events would have been very different if

small events had worked out differently; see, for example, Handlin, Oscar, Choice or Destiny: Turning Points in American History, Atlantic Monthly Press, Boston, Mass., 1954. However, as the last two chapters of this example make clear, it is usually not considered acceptable to imagine just how the subsequent events might have been different, concretely.

(37) It has been suggested that differences in behavior of "high" and "low" situations may make such a treatment valuable, or that an actual division of the world into "high" and "low" makes it desirable. However, just because our sample is weak in the middle range of the spectrum, we do not have to conclude that we have to break our sample into smaller subsamples, capable of supporting less detailed analysis. If there is some qualitative difference in behavior in the different zones, this difference in behavior may be tied to a smooth continuum of different behaviors, as one moves from one pole to the other. Even if a clearcut threshold effect does exist, then, in order to explain this effect, operating on a set of continuous variables, we would normally study the discontinuous implications of the continuous interactions of the original continuous variables. Exceptions may exist, but, in more cases than one might expect apriori, it is better to treat continuous variables as such, even if they have strange properties.

(38) Strictly speaking, it would be more accurate to say that verbal or mathematical models derived from external empirical data alone do not incorporate the information, both quantitative and structural, to be derived from accounting for the mutual underlying resemblance of different human brains. Once one's "empathy" has led one to postulate a certain model structure, one can, of course, try to translate this model into a related empirical model for empirical estimation; even then, however, the empirical test would only account for one of two sources of validation.

(39) Von Neumann, John and Morgenstern, Oskar, The Theory of Games and Economic Behavior, Princeton U. Press, Princeton, NJ, 1953, p.15-33.

(40) See note 1.

(41) Flavell, John H., The Developmental Psychology of Piaget (including Foreword by Piaget), Van Nostrand, Princeton, NJ, 1963.

(42) The most popular reference would be Erikson, Erik H., Insight and Responsibility, Norton, NY, 1964, p.111-134. The school of ego psychiatrists is much larger, but less mapped out than many other fields for the wandering political scientist; the concept of stages, while often present, often requires digging out. Another reference, less transparent but also influential: Hartmann, Heinz, Ego Psychology and the Problem of Adaptation, International Universities Press, NY, translated by D. Rapoport, 1958. One advantage of these approaches is that they can be more easily compared with cybernetic views, emphasizing human intelligence.

(43) One must make allowance for a few state parameters, however - such as metabolic, respiratory and hormone levels - which are less often subject to learning. Sex and intelligence may both be affected by such variables. However, to say that learning may proceed faster or slower does not invalidate a person's ability to learn, in most cases. On the behavioral level, flexibility remains critical, particularly when we are speaking of heads of state and the like, who are rarely literal imbeciles.

(44) The reference from Triffin, in note 45, curiously enough, implies quite strongly that "our textbooks" have emphasized these points about perfect competition. Ferguson, C.E., Microeconomic Theory, Irwin Series in Economics, Illinois, 1969, Revised Edition, Third Printing: methodological introduction refers to the "extreme aprioristic" school of microeconomists, "prominent since John Stuart Mill"; much of the rest of this text deals with the classic theory of perfect competition and its later developments.

(45) Triffin, Robert, Monopolistic Competition and General Equilibrium Theory, Harvard University Press, Cambridge, Mass., 1956, p.5: "For most of Professor Chamberlain's and Mrs. Robinson's readers, this is the basic distinction between

monopolistic (or imperfect), and pure (or perfect) competition. If the sales curve of the firm is perfectly elastic, we are concerned with pure competition. If, on the contrary, the curve is tipped, competition is taken to be monopolistic or imperfect... The substitution of the equation of marginal cost and marginal revenue for the less general and less elegant equation of marginal cost and price has been the main contribution of monopolistic competition theory to the 'pure economics' of our textbooks."

(46) Ferguson, C.E., op.cit., Chapter 14.

(47) Lindblom, Charles E., "The Rediscovery of the Market", especially p.441, in Readings in Economics, Paul A. Samuelson, ed., McGraw-Hill, NY, 1970, Sixth Edition.

(48) Lindblom's article, note 47, provides some evidence on this point. In the fall of 1973, an article appeared in the New York Times indicating that more than 90% of Soviet consumer industries, at a minimum, had been converted to the Lieberman system. The New York Times Index at this writing was complete up to August 15, 1973; it listed a major article on the front page, June 3, 1973, elaborating on how thoroughly the conversion has been made, and at any rate quoting Pravda on Soviet condemnation of those who oppose the new methods. While the same Pravda article was discussed briefly on June 6, p. 23, in the Washington Post, we were unable to find the Times article in its indexed location at Harvard, or in nearby locations that we looked at. However, subsequent copies of the Index should clarify these points.

(49) See Samuelson, Paul A., Foundations of Economic Analysis, Harvard University Press, Cambridge, Mass., 1947. Lagrange multipliers are used in maximizing a fixed function, subject to static constraints. These multipliers correspond, essentially, to prices.

(50) Wiener, Norbert, Cybernetics, or Control and Communication in the Animal and the Machine, MIT Press, Cambridge, Mass., 1961, Second Edition.

(51) Wiener, Norbert,"Perspectives in Cybernetics," in Wiener, Norbert and Schade, J. P., eds., Cybernetics of the Nervous System, Elsevier Publishing Co., Amsterdam, Ny, 1965. In this article, Wiener emphasizes the statistical subdivision of cybernetics, in particular, as the portion of cybernetics most worth pursuing. This is quite close to our suggestions here, insofar as the mathematics of Chapter (II) tend to be part of the statistical subdivision, but Wiener's specific suggestions are very different from our own, in terms of the overall explanations they point to for gross behavior; Wiener emphasizes patterns of resonance among multiple sources of radiation, an idea which could conceivably relate to our technique of pattern analysis, but the connections do not appear simple. Wiener goes on to suggest strongly that the new statistics developed to deal with the analysis of time-series (one-dimensional phenomena) for living systems may someday be extended to statistical physics (four-dimensional phenomena, thermodynamics) and provide a revolutionary new understanding of the possibilities for maintaining order in equilibrium.

(52) Deutsch, Karl W., The Nerves of Government, Free Press, Glencoe, 1966. p.xxvi: "In the main, these pages offer notions, propositions and models derived from the philosophy of science, and specifically from the theory of communication and control - often called by Norbert Wiener's term "cybernetics" - in the hope that these may prove relevant to the study of politics, and suggestive and useful in the essential development of political theory that will be more adequate - or less inadequate - to the problems of the later decades of the twentieth century."

(53) A few brief hints may be in order here, to indicate the existence of specific possibilities. If one presumes that some sort of inborn "reinforcement" mechanism provides the brain with a current measure of a cardinal utility function to maximize, then section (x) of Chapter (II) indicates the optimal way to adapt an elaborate behavior-generating network, to maximize this function, conditional upon the availability of a network model of the "external" environment. This involves the passing back of "ordered

derivatives", a quantitative piece of information, represented by some physical information flowing backwards along the same network which overtly carries only gross, direct behavior-generating information (electrical impulses); the microtubules, which criss-cross almost all neurons, could well be implementing a hidden network of chemical feedback of this kind, carried back from cell to cell to cell, originating in the hypothalamus and epithalamus (effectively in the pituitary and pineal, whose exact rules of operation we would not pretend to know at this point.). Network models to predict the external environment, as in Chapter (II), could be generated ("estimated") by a similar mechanism, based on measuring predictive accuracy at some sites like the glomeruli of the thalamus. (This hypothesis yields the empirical prediction that certain states of chronic insensitivity and rigidity in behavior governed by the cerebrum would be replaced by normal cerebral learning, if only the inputs could get as far as the glomeruli; this could be accomplished either by nerve growth factors, synthesized to enhance the growth of random connections from the hypothalamus or epithalamus to the glomeruli, or even by learning procedures which take full advantage of the microscopic bootstrap process which develops new connections to the glomeruli under normal conditions. This prediction is not only testable, but also of potential practical value.) The giant pyramids of the cerebral cortex might be performing pattern analysis, as in section (ix) of Chapter (II); the duality of the functions f-sub-i and g-sub-i, in generating and predicting the same pattern-description variable, may well correpond to the dual poles of these cells. The time factor, of course, requires that all of these ideas only be approximations; still, they may have some suggestive value even in experimentation.

(54) See the discussion by Herbert Simon on "optimizing" and "satisficing," in Lazarsfeld, Paul F., ed., Mathematical Thinking in the Social Sciences, Free Press, Glencoe, Ill., 1954. A deeper understanding of Simon's observations would require, of course, a more general framework.