CONTROL THEORY AND INSURANCE SYSTEMS

by

ALEXANDROS A. ZIMBIDIS

Thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

THE CITY UNIVERSITY, LONDON
DEPARTMENT OF ACTUARIAL SCIENCE AND STATISTICS

APRIL 1999

Contents

Chapter 4 Application to Insurance Pricing

Appendices

Acknowledgements

I would like to express my sincere thanks to Professor Steven Haberman, Dean of the School of Mathematics, Actuarial Science and Statistics for his invaluable help and guidance as my supervisor in this thesis.

I am also grateful to all my family and especially to my wife Nikoletta for her patience and support during this project.

Finally, I would like to dedicate the whole work to my sons

Anastasios and Joseph.

Declaration

Abstract

The primary target of the current thesis is the establishment of a link (bridge) between the abstract concepts of control theory and the practical applications connected with different insurance systems. The existence of such a link has been identified by several authors in the last two decades who have focused on the use of conventional control theory (and optimal control techniques) in actuarial problems.

In our approach, after providing a short reference guide to the modern control theory in chapter 2 and a selective review of the literature in chapter 3, we propose three distinct models covering the most important areas of insurance applications (i.e. Life, General Insurance & Pensions). From the control point of view we actually use all the potential tools i.e. Multiple Input - Multiple Output, Time-Varying format, Non-linear equations, Stability Analysis, Root Locus Method, Optimal design for the parameters involved and optimal control of a dynamic system.

The thesis deals with the basic concept of an insurance system, "the premium" and aims to answer the critical question "How to calculate and control the premium rating process?".

In the first model (chapter 4), we examine the general process of insurance pricing, using the standard equation which connects the three major variables involved i.e. premiums, claims & surplus. Starting from the roots of actuarial science and the static point of view and passing to Lundberg's revolution with his dynamic view, we arrive at the modern alternative view of control theory with respect to pricing models. We concentrate to the concept of stability rather than to the traditional concept of ruin.

In the second model (chapter 5), we construct a dynamic system which describes a special reinsurance arrangement (multinational pooling) which may only be handled using the modern control theory (as it refers to a multivariable system). Actually it is an extension of the previous model to a multi-system which consists of different subsystems, as described in chapter 4 considering also the interaction between them.

Finally in the third model (chapter 6) we investigate the philosophy and mechanisms of the Social Security System and the PAYG funding method. It may also be seen as an extension of the first model to the area of pensions and the necessity of calculating and controlling the respective contribution rate and the age of normal retirement. At the end of the sixth chapter, a simulation is carried out for the projected population data of Greece up to the year 2020.

Chapter 1

Introduction

1.1 The insurance Paradox!

"Anything reasonable is correct!"

Unfortunately... or fortunately the above sentence is not always true!

It is extremely difficult to establish what is the real meaning of "reasonable" or "logical" in order to evaluate the truth or false of the first sentence.

Anyway, we shall not go further with philosophical questions. Simply we could say that, there are some complicated problems in different scientific areas wrapped in the colourful paper of easiness.

Of course, the most interesting thing arises when these problems come out from the real world and they are not just theoretical creatures of a scientific laboratory or clever mathematical puzzles.

Hilary Seal has pointed out such a strange ("complicated-obvious") insurance problem. In a paper which he had delivered to a conference on simulation organized by the Research Committee of the Society of Actuaries (1970), he set up a simple model for the total annual claims of a motor insurer and proposed several "reasonable" rules for setting the premium rates based on recent experience. (e.g. the average of recent years in the days before inflation).

The surprising result of his research was that nearly all the simulated companies went broke.

At that moment there was again a simple answer. "Don't worry all those mathematical formulae and simulation procedures are not reliable. Insurance companies use the average function all previous years and they are doing well up to now".

Of course there was a smarter action... to go on with further research on the specific topic and that was done afterwards by S. Benjamin who attended that conference and found some first answers to his earlier questions.

At that time, S. Benjamin had been involved, giving advice on the control of a certain type of a new non-life insurance portfolio. He was faced with the same "obvious" problem of establishing a strategy to control an insurance system.

1.2 The successful marriage

The problem of the last section was not a new one. Actuaries have been working on financial controlling methods of (especially) insurance systems over the last 200 years. They have developed different practices to manage long or short-term insurance funds and in more recent times pension funds.

S. Benjamin and L.A. Balzer produced some papers (and a joint one) and were amongst the first people who realized that insurance problems may be placed in a wider theoretical context named Control or Dynamic System Theory.

Control Theory, a very promising theory has been "married" with different partners over the last decades engineering (the first one), physics, Aeronautics, chemistry and economics.

The very last partners are finance and actuarial science. There are a lot of people who think that the last marriage will be one of the best and most successful of recent years.

1.3 A short reference to the traditional actuarial work

In order to understand better why the marriage mentioned in the last section may be proved absolutely successful, we should consider a short reference to the traditional actuarial work up to a few years ago.

Actuarial Science has been based on the statistical point of view using also some other elementary methodologies from other mathematical areas. Of course a special way of thinking has quickly appeared among the actuarial practitioners or theorists which (way of thinking) has been retained within a closed group of "Fellows".

Now tracing the development of traditional actuarial work we may consider the following stages:

(1) Pre-historical age: At that stage nothing formal has been established but practitioners involved with commercial activities may well use the simple average function in order to evaluate risks. That stage may go back even to the origins of the humanity.

(2) Statistical point of view: Now reaching the roots of historical period we may find efforts from different people (normally with some kind of mathematical background) in order to establish more than the simple average function for each risk. Concepts as the variance, standard deviation or other moments of higher degree, reserves mortality tables etc. have been proposed. All that work is described as individual risk theory providing only the static picture of a problem (like a photograph).

(3) Dynamic point of view: It was just the first decade of the 20th century in which Lundberg attempted (successfully at that time period) to upgrade the static point of view into a dynamic one. He considered an insurance system over an extended time period (of more than one unit time) and focused on the calculation of probability of ruin.

Lundberg may be considered as the man closing the traditional scene of actuarial science while at the same time writing the prelude of a new concept. More analysis about the traditional actuarial approach (also providing some formulae) will be given in sections (4.4) and (4.5).

A complete history for the actuarial science and profession is provided in Bühlman (1997). Starting from the establishment of Equitable Life Assurance Society in 1762 (upon certain traditional techniques and statistical data) the author traces all the steps (as listed below) up to the modern view of risk, the advanced mathematics of 20th century and the challenge (for the actuaries) from the financial world.

- Establishment of the first life insurance companies in the second half of the 18th century.
- Formation of the first national actuarial bodies (the Institute of Actuaries in 1848) in the second half of the 19th century.
- The first International Congress of Actuaries in Brussels in 1895.
- The high recognition for the actuaries as the persons who controlled the know-how of the life insurance industry in the first half of the 20th century ("golden age").
- The great dispute for the actuaries both for their professional and academic contribution in 1950's.

- Actuarial techniques began to break out from the area of life insurance in 1960's and 1970's entering the field of general insurance (and pension funds).
- In more recent times, actuaries (and academics) compete equally with other professions to provide their services and expertise in financial markets and business world.

1.4 The roots of a new direction for actuarial science

After Lundberg's prelude, and due to the tendency which appeared early this century of crossing the boarders of different scientific areas, actuarial research has been affected rapidly from various directions. Economics, Pure and Applied Mathematics offered some very promising alternatives for actuarial science in order to enhance and make more powerful its ability of exploring insurance problems. For example utility theory offered an entirely new and unique view for pricing and handling risks. But the most powerful proposal arose from pure mathematics, the theory of dynamic systems or qualitative analysis of differential / difference system of equations or control theory.

If we want to draw an abstract line from Lundberg (1909) who introduced the dynamic view for insurance problems up to the recent years we may have the following names. Lundberg, De Finetti, Borch, Balzer, Benjamin, Martin-Löf, Rantala, Taylor, Vanderbroeck, Loades, Haberman. Control theory examines an insurance system not only from the ruin point of view but provides deep insight into the mechanisms of the information or decision making system. It helps us to avoid enormous amounts of simulations by considering the quality characteristics of a system. Further analysis of the research work of each name mentioned in this section will be carried out in Chapter 3.

1.5 Description of thesis structure & its objectives

Closing the introductory chapter, we shall present the general structure of the thesis with a brief description of each subsequent chapter regarding the topic examined and the objective aimed to.

Chapter 2: It stands as an introductory and quick reference point for all the basic concepts and results of control theory. Accompanied with the respective appendices it is self-contained with respect to the mathematical background. A simple insurance system is taken as an example in order to apply the concepts and results of control theory. This kind of development, facilitates the better and deeper understanding while also points the links between the two scientific areas.

Chapter 3: A literature review is considered with description and critical comments provided for each paper. Starting from De Finetti (1957) and following the footprints of the research for the last 30 years we are trying to reach the highway of a new approach for insurance problems.

Chapter 4: The first application of control theory is implemented regarding the problem of "Insurance Pricing". Firstly we provide the traditional solution (with some other variations). Secondly we present the entirely new approach using control theory. The current research actually extends the work of Benjamin and Balzer (1980) incorporating design concepts from other papers and also some new ones. The actual problem refers to a non-life insurance portfolio. From the control point of view the model is considered to be a single input - single output one with two versions

(a) time invariant and (b) time varying format

Chapter 5: A more complicated model is considered in chapter 5 as we design a multiple input - multiple output system. This problem requires (unavoidably) the use of modern control theory. The abstract system refers to special reinsurance arrangements similar to multinational pooling. Of course the abstract system may be also applicable to other practical problems as capital allocation between subsidiaries companies or solbency margin requirements between different lines of business.

Chapter 6: In order to show the wide use of control theory to all insurance problems, an application for PAYG funding method is designed in chapter 6. The original problem and the respective system is non-linear and is being handled by the use of optimal control theory. At the end of this chapter we consider a special application for the Greek population using some projected data up to the year 2020.

Chapter 7: Conclusions, important results and scope for further research is presented at the final chapter.

1.6. Notation of the thesis

Developing the current thesis we have attempted to use a uniform notation in all chapters either describing past papers or expressing our own ideas. We have also tried to keep unique symbols for all the parameters but that has been achieved only for the important ones, as the large volume of parameters and special values forced us to make some duplications (or move repetitions). Here we shall refer to three sections (4.3), (5.3) & (6.10) which contain the important symbols used almost through the whole thesis while also state that all the other symbols (not mentioned there) have limited use in each chapter or section.

We should also refer to the capital letters A,B,C,D which represent the basic matrices of the standard format of a dynamic system (see equations (2.6.3) & (2.6.4)) while $\underline{u}_n$, $\underline{x}_n$, $\underline{y}_n$ represent the input, state and output vector variables respectively. Finally, state that the subscript (n) declares a discrete model while the (t) variable declares a continuous model.

Chapter 2

Control Theory Concepts

2.1 Introduction

It is very important and extremely useful (not only for the need of this thesis) to summarize the basic concepts and theorems of control theory, providing critical comments and examples concerning the translation of these ideas into the actuarial problems.

In this chapter, we consider a very simple example (actually the Balzer - Benjamin's model) of group profit sharing schemes. Then, we try to go through the basic concepts of control theory, considering at each section how this certain idea helps our modeling or how it facilitates a new further insight into the problem.

Some of this further insight had been already pointed out and examined from past papers of Balzer and Benjamin (1980) using traditional control techniques. Our effort is to emphasize the approach of modern control theory (using vectors, matrices, eigenvalues, eigenvectors, etc...).

We think this approach is more powerful (and the only way) to handle efficiently the systems of multiple input and output. We are also trying to examine and translate properly the concept of controllability and observability which has not been considered before.

Finally, our aim (and perhaps ambition) in this chapter is to provide a complete introductory guide to control theory for actuarial scientists. (Consequently we quote some results which are not used directly in the technical development of the models but which complement the whole discussion).

2.2 A simple (standard) actuarial problem

Profit - sharing schemes are widely used not only in group life policies but also in non-life policies and in large reinsurance portfolios. Here, we consider a group life policy which is effected at time n=0 with a mutual agreement of the insurer and the policy holder upon the following items.

(1) The initial premium P_1 paid for the first year of policy.

(2) The required expense and profit margin of insurer 1-e (expressed as a percentage of gross premium, so actually $e \cdot P_n$ is used to cover the claim cost for each year n, where P_n is defined below).

(3) The modification of each renewal premium for year n according to the most recently known experience of year n-2 (so the risk premium will be based on C_{n-2}) and the accumulated surplus up to that year (a proportion ε of accumulated surplus will be refunded). Hence, we have the formula

$$P_n = \frac{1}{e} \cdot C_{n-2} - \varepsilon \cdot S_{n-2} \quad , \quad n=2,3,... \tag{2.2.1}$$

(C_n and S_n are defined below).

(4) Accumulated surplus is calculated using the following formula (at the end of each year).

$$S_n = S_{n-1} + e \cdot P_n - C_n \quad , \quad n=1,2,... \tag{2.2.2}$$

where $S_0 = P_0 = C_0 = 0$ (2.2.3)

(5) No interest or inflation factors are considered.

(6) P_n stands for the premium received in year n (n=1,2,...)

S_n stands for the accumulated surplus at time n (n=1,2,...)

C_n stands for the claims incurred in year n (n=1,2,...).

In the next sections, we shall formulate the problem using modern control theory. The consideration of the certain model will secure direct comparison between our results and those of Balzer and Benjamin's.

2.3 Physical & other types of systems -
- Modelling & Design - Mathematical Representation

The real world, we live in, is full of different types of physical systems which affect our lives, work, behaviour and generally all human activities. Even the whole universe could be considered as a large (perhaps the largest) system divided into smaller subsystems.

Common sense describes the word "system" as a set of different elements which may be related together (in some way). A well established system should have a well established set of elements and well defined relationships between them. Quoting the system's definition of Ogata (1970) "A system is a combination of components that act together and perform a certain objective".

Examples of physical systems may be an electric circuit, a group of galaxies in the universe, the plants of a tropical island. Of course the systems are not limited to physical ones so other examples may be: The whole economy of a country, the student society of a University, the stock exchange, the insurance market, etc... (We stop here, otherwise we could end up our thesis just describing the different types of systems in the real world).

Although it may appear strange, all the above systems have a common point! They all can be translated and fully represented with a set of mathematical equations. Of course there is a certain way (a bridge) which transfers the elements and relationships of a system into numbers and equations.

The bridge (mentioned above) is the modeling procedure of our system. A successful model can give us more than half of the final solution. We should stress that in most times there are many ways of modeling and designing a certain system.

The first step of any dynamic analysis is to describe clearly the process and its unique characteristics. Then establish a system of differential (or difference for discrete type of systems) equations which is the mathematical representation of that verbal description. The system is considered to be dynamic since all the parameters are functions of the time variable n.

In our example (section (2.2)) we have already translated the insurance system and we have produce its mathematical representation as follows:

$$S_{n+1} = S_n + e \cdot P_{n+1} - C_{n+1} \tag{2.3.1}$$

$$P_{n+1} = \frac{1}{e} \cdot C_{n-1} - \varepsilon \cdot S_{n-1} \tag{2.3.2}$$

It is obvious from the structure of equations (2.3.1) and (2.3.2) that we refer to a linear dynamic system (as all the variables depend upon time).

2.4 Well Oriented Systems (Input - Process - Output)

Although it appears that all systems could be described as well oriented systems (simply that means, there is a type of order) this is not always true! There are some not well-oriented systems especially in quantum mechanics where the order of cause and effect is usually distorted.

At the present moment the insurance systems (which is our final target) do not obey the laws of quantum mechanics. So, we shall continue and focus on well-oriented systems which have an inertia order and this order may be generally described with three words: input - process - output.

Now we shall return to our example determining the variables of our problem and connect each one with some of the three words above.

A first trial to order our variables will be the following

premiums - claims - surplus

Thinking by common sense, premium is the input of the system (because it comes first), claims represents the process (think of the actuarial wording usually used "the claim process") and finally surplus is the output (because it comes as the final result of the whole procedure).

The approach above, may be a way of thinking but it is not the correct way of modeling the variables using control theory techniques. If we think deeper we shall find that the right order is the following:

claims - surplus - premium

or diagramatically,

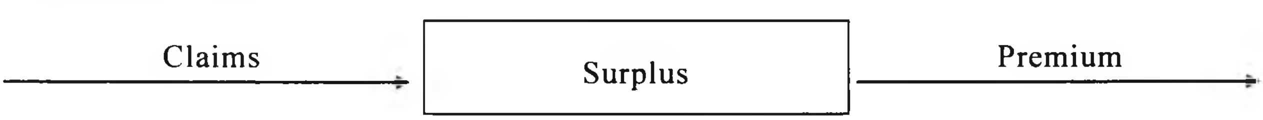


Diagram (2.4.1)

As it was described in the development of the model (section (2.2)) there was an initial agreement for the first year's premium and consequently the system was free to adjust the renewal premiums of each subsequent year according to the claim experience and accumulated surplus. Hence, the claim variable comes first (as input), modifying the surplus variable (assuming a certain process) and finally producing the output of the system i.e. the next year's premium.

As we can see, there is a mechanism which coordinates the different variables of the system. This kind of mechanism is called feedback action. Actually at the end of each time period a certain portion of the surplus is being fed back to the system in order to be controlled. So accumulated surplus may be defined as the measuring element in the whole process indicating the state of the system.

2.5 Control theory diagrams

The control diagrams have similarities and differences with the usual graphs or other types of diagrams. Blocks, arrows, cycles and signs (+,-) are used for the representation of the basic parameters of the system. Typical examples are given below: Diagrams (2.5.1), (2.5.2), (2.5.3).

i) The first one is an open loop system. Diagram (2.5.1).

ii) The second one is a closed loop system with a feedback control mechanism Diagram (2.5.2).

iii) The third one is again a closed loop system with a feedforward control mechanism Diagram (2.5.3).

Control diagrams are very useful as a geometrical representation of the process of the system.

Diagram (2.5.1)

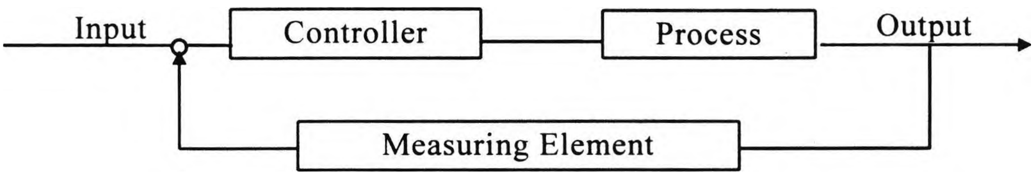


Diagram (2.5.2)

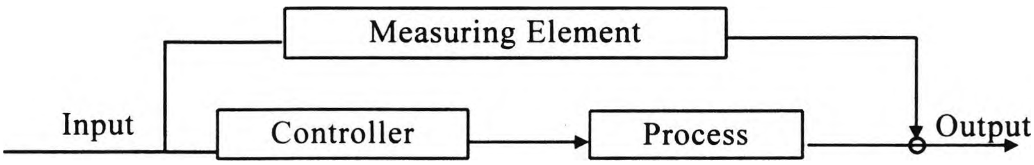


Diagram (2.5.3)

The diagrams above are similar with the ones quoted in Benjamin (1984).

At this point it will be interesting to show the relationship between a simple open-loop system with the respective closed-loop one (assuming a feedback mechanism).

Consider the open-loop system

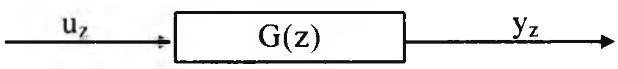


Diagram (2.5.4)

where G(z) is the transfer function and

u_z is the z-transformed input while

y_z is the z-transformed output

i.e. $y_z = G(z) \cdot u_z$

Z transformation is discussed in Appendix I.

then if we introduce a negative feedback mechanism with H(z) as a transfer function (see the next diagram (2.5.5)).

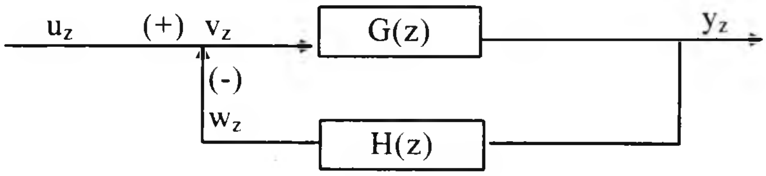


Diagram (2.5.5)

we shall obtain that

$$y_z = \frac{G(z)}{1+G(z)H(z)} \cdot u_z \tag{2.5.1}$$

hence the transfer function of the closed loop system will be

$$\frac{G(z)}{1+G(z)H(z)}$$

Finally we shall provide the required justification for the relations above.

From the diagram (2.5.5) we obtain the relationships:

$$y_z = G(z) \cdot v_z \tag{2.5.2}$$

$$v_z = u_z - w_z \tag{2.5.3}$$

$$w_z = H(z) \cdot y_z \tag{2.5.4}$$

Substituting (2.5.4) into (2.5.3) we obtain

$$v_z = u_z - H(z) \cdot y_z \tag{2.5.5}$$

and now solving the system of (2.5.2) and (2.5.5) we find that

$$y_z = G(z) \cdot [u_z - H(z) \cdot y_z] \Leftrightarrow$$

$$(1 + G(z) \cdot H(z)) \cdot y_z = G(z)\, u_z$$

and from the last relationship (rearranging the terms) we obtain (2.5.1)

2.6 Input - State - Output Spaces

The powerful tool of the modern control theory is the state-space analysis of the system. Quoting from Ogata (1970) the necessary definitions for this kind of analysis we have the following:

Definition (2.6.1) "State: The state of a dynamic system is the smallest set of variables (called state variables) such that the knowledge of these variables at $n=n_0$ together with the input for $n \geq n_0$, completely determines the behavior of the system for any time $n \geq n_0$".

Definition (2.6.2) "State variables: The state variabes of a dynamic system are the smallest set of variables which determine the state of the dynamic system".

Definition (2.6.3) "State Vector: If h state variables are needed to completely describe the behavior of a given system then these h state variables can be considered to be the h components of a vector $\underline{x}_n$. at time n. Such a vector is called state vector.

Hence, generally speaking in every system we may distinguish three critical spaces: (The symbol $\cong$ stands for equivalent between two vector spaces)

(1) $U \cong \mathbb{R}^{\ell}$ (input space): the vector space where the input variable $\underline{u}_n$ can take its values.

(2) $X \cong \mathbb{R}^{h}$ (state space): the vector space where the state variable $\underline{x}_n$ can take its values.

(3) $Y \cong \mathbb{R}^{m}$ (output space): the vector space where the output variable $\underline{y}_n$ can take its values.

where $\mathbb{R}$ is the space of real numbers.

Rewriting equations (2.3.1), (2.3.2) and substituting the second into the first we obtain

$$S_{n+1} = S_n - \varepsilon e S_{n-1} + C_{n-1} - C_{n+1} \tag{2.6.1}$$

$$P_{n+1} = -\varepsilon \cdot S_{n-1} + \frac{1}{e} C_{n-1} \tag{2.6.2}$$

The format above is similar to the standard format of the mathematical representation of a discrete dynamic system i.e.

$$S_d(A,B,C,D): \begin{cases} \underline{x}_{n+1} = A \cdot \underline{x}_n + B \cdot \underline{u}_n & \text{where } A \in \mathbb{R}^{h \times h}, B \in \mathbb{R}^{n \times \ell} \quad (2.6.3) \\ \underline{y}_n = C \cdot \underline{x}_n + D \cdot \underline{u}_n & \text{where } C \in \mathbb{R}^{m \times h}, D \in \mathbb{R}^{m \times \ell} \quad (2.6.4) \end{cases}$$

the standard format for continuous type is the following:

$$S_c(A,B,C,D) \cdot \begin{cases} \dot{\underline{x}}(t) = A \cdot \underline{x}(t) + B\underline{u}(t) & (2.6.5) \\ \underline{y}(t) = C \cdot \underline{x}(t) + D\underline{u}(t) & (2.6.6) \end{cases}$$

where A,B,C,D defined as before and $\dot{\underline{x}}(t)$ is the first derivative of $\underline{x}(t)$. (In the continuous form we use t as the time variable instead of n.

From this point we shall use S(A.B.C.D) when a concept is applicable either for a discrete or a continuous dynamic system.

As we observe the A,B,C.D matrices are constants. (i.e. do not depend on the time variable t or n). These systems are called time invariant. Of course, physical systems may be non-time invariant and consequently the matrices depend on the time variable i.e. A(t), B(t), C(t), D(t) or A_n, B_n, C_n, D_n.

The graphical representation of systems (2.6.3) and (2.6.4) is given in diagram (2.6.1).

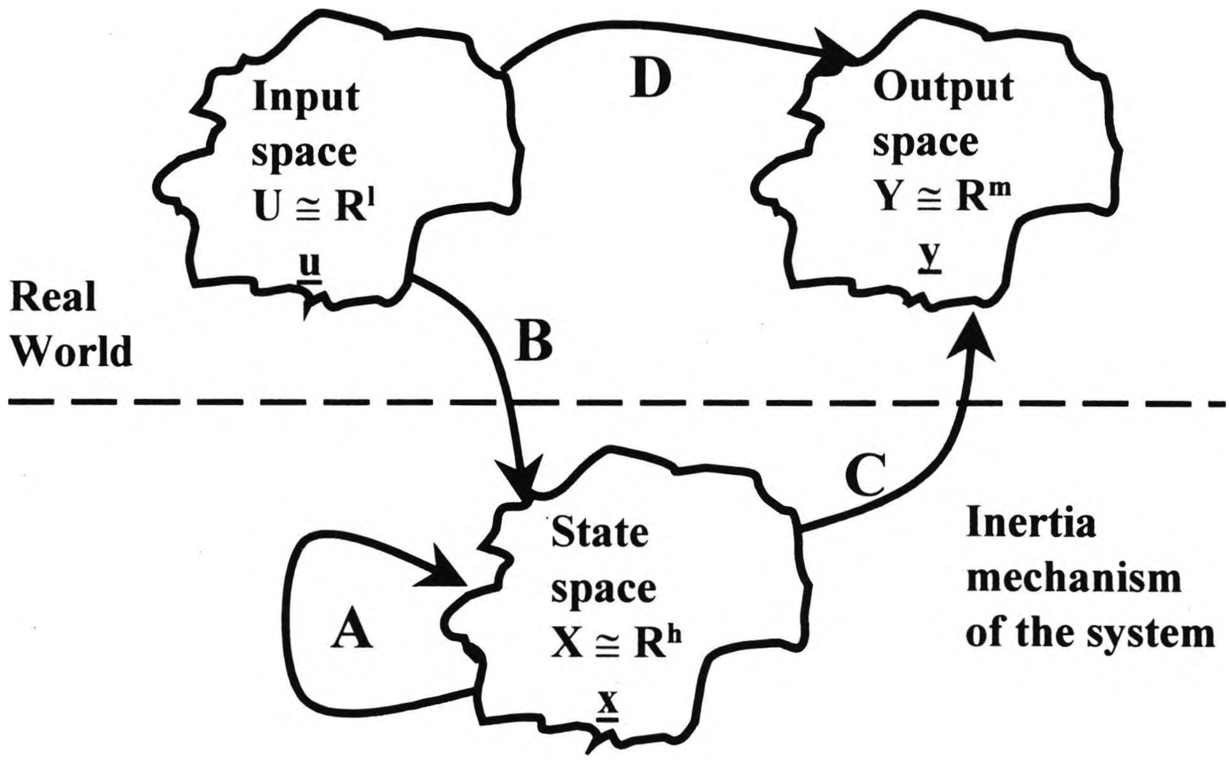


Diagram (2.6.1)

In order to find the input, state and output vectors of our example (and consequently the respective spaces) we must find the standard vector format of the system. This will be done in the next section and the immediate results will be the following:

$$U \cong \mathbb{R}^3,\ X \cong \mathbb{R}^2,\ Y \cong \mathbb{R}$$

2.7 Standard vector format of our model

As we have seen in the last section, our problem can take a similar format as the typical mathematical representation of a dynamic system. Here, we shall try to convert the equations into the vector form and achieve the exact standard format.

Let $\underline{x}_n = \begin{bmatrix} S_n \\ S_{n-1} \end{bmatrix},\ \underline{u}_n = \begin{bmatrix} C_n \\ C_{n-1} \\ C_{n-2} \end{bmatrix},\ \underline{y}_n = [P_n]$

with zero initial conditions $\underline{x}_0 = \underline{u}_0 = \underline{y}_0 = \underline{0}$

Then the system of equation (2.6.1) and (2.6.2) is written in the form,

$$\begin{bmatrix} S_{n+1} \\ S_n \end{bmatrix} = \begin{bmatrix} 1 & -e\varepsilon \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} S_n \\ S_{n-1} \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} C_{n+1} \\ C_n \\ C_{n-1} \end{bmatrix} \tag{2.7.1}$$

$$P_{n+1} = \begin{bmatrix} 0 & -\varepsilon \end{bmatrix} \begin{bmatrix} S_n \\ S_{n-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{e} \end{bmatrix} \begin{bmatrix} C_{n+1} \\ C_n \\ C_{n-1} \end{bmatrix} \tag{2.7.2}$$

or equivalently,

$$\underline{x}_{n+1} = A \cdot \underline{x}_n + B \cdot \underline{u}_{n+1} \tag{2.7.3}$$

$$\underline{y}_{n+1} = C \cdot \underline{x}_n + D \cdot \underline{u}_{n+1} \tag{2.7.4}$$

$$A = \begin{bmatrix} 1 & -e\varepsilon \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & -\varepsilon \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 & \frac{1}{e} \end{bmatrix}$$

Having, achieved the standard format of mathematical representation, we can go through all the basic concepts of control theory and apply them into the certain problem (providing the required verbal interpretation of each idea).

At this point we should stress again, the problem could be handled by traditional methods, i.e.

(1) Using difference equation theory (see Goldberg (1963))

(2) Traditional Control Theory (see Balzer and Benjamin (1980)).

The vector representation is more powerful and gives us further insight into the problem. It also facilitates a quick and simple application of the basic theorems of control theory and explain the critical properties of controllability, observability and stability of the system.

Actually the insight into the problem is gained by comparing the vector format of our example with the standard format. For example it is obvious now that the state of the system is described from S_{n-1} and S_n and not only from S_n i.e. We need the knowledge of both S_{n-1} and S_n (or the vector $\underline{x}_n$) along with the input in order to completely determine the future behavior of the system.

Further insight is also gained because the description and equations have been replaced by a set of four matrices A,B,C,D. So we actually face the problem $S_d(A,B,C,D)$ and our research is based on the exploration of the properties of these (4) four items. We shall see in the next sections how we can obtain immediate results from the size and pattern of the matrices.

2.8 Solution of the general form of a discrete dynamic system

Let us consider the general form of a discrete dynamic system

$$\left.\begin{aligned} \underline{x}_{n+1} &= A\cdot\underline{x}_n + B\cdot\underline{u}_n \\ \underline{y}_n &= C\cdot\underline{x}_n + D\cdot\underline{u}_n \end{aligned}\right\} S_d(A,B,C,D)$$

We may easily obtain its solution by mathematical induction calculating the first equation for values n=0,1,2,... Let $\underline{x}_0, \underline{u}_0$ the initial conditions, then from the first equation of the system above we obtain,

$n=0,\ \underline{x}_1 = A\cdot\underline{x}_0 + B\cdot\underline{u}_0$

$n=1,\ \underline{x}_2 = A\cdot\underline{x}_1 + B\cdot\underline{u}_1 =$

$= A(A\underline{x}_0 + B\cdot\underline{u}_0) + B\underline{u}_1 =$

$= A^2\cdot\underline{x}_0 + (AB\cdot\underline{u}_0 + B\underline{u}_1)$

$n=2,\ \underline{x}_3 = A\cdot\underline{x}_2 + B\underline{u}_2 =$

$= A(A^2\cdot\underline{x}_0 + (AB\underline{u}_0 + B\underline{u}_1)) + B\underline{u}_2 =$

$= A^3\cdot\underline{x}_0 + (A^2B\underline{u}_0 + AB\underline{u}_1 + B\underline{u}_2)$

$\vdots$

It is easily proved (mathematical induction) that

$$\underline{x}_n = A^n\underline{x}_0 + \sum_{k=0}^{n-1} A^k B\underline{u}_{n-k-1} \quad , n=1,2,... \tag{2.8.1}$$

We shall proceed with the solution of the 2nd equation of system $S_d(A,B,C,D)$.

$n=0, \qquad \underline{y}_0 = C\cdot\underline{x}_0 + D\cdot\underline{u}_0$

It is easily proved directly by substitution of (2.18.1) into the 2nd equation of system $S_d(A,B,C,D)$ that,

$$\underline{y}_n = CA^n\underline{x}_0 + CA^{n-1}B\underline{u}_0 + \cdots + CB\underline{u}_{n-1} + D\underline{u}_n, \quad n=1,2,... \tag{2.8.2}$$

If we denote $\Phi(n)=A^n$, it is obvious that the solution is based on this specific matrix-function. The matrix A^n is called fundamental matrix since the resultant response of the system is fundamentally based on it.

The matrix function $\Phi(n)$ is also called state-transition function (matrix) and for the time-varying systems is of the form $\Phi(n,n_0)$ i.e. depending on initial conditions too.

In order to complete the analysis for the time-invariant systems we shall only refer to the respective state transition matrix of the continuous form i.e.

$$\Phi(t) = e^{At} = \sum_{k=0}^{\infty} A^k \frac{t^k}{k!}$$

The discussion above is based on Gadzow (1973) and Ogata (1970).

2.9 Application of the general solution to our model (Calculation methods for the fundamental matrix)

It is obvious from the last section that we may immediately obtain the solution for our problem using the general form of the solution and calculating the powers of the square matrix $A \in \mathbb{R}^{h \times h}$ (powers for n=2,3,...). Here we shall see two methods the direct and the indirect one (for the calculation of A^n, n=2,3,...).

There is an extensive theory of linear algebra for calculating powers of a square matrix A, or generally calculating functions of A. (As we have seen the expression e^{At} is needed for calculating the general solution of a continuous linear time invariant dynamic system). A brief introduction of the theory above is given in Appendix II. Assuming the required study of appendix II, we shall proceed with the calculation of the n^{th}-power of A without detailed explanation.

Generally, A^n can be obtained as a product of three other matrices i.e.

$$A^n = Q \cdot J^n \cdot Q^{-1}$$

where Q a matrix produced by the eigenvectors of A (or generalized eigen vectors) and J is a matrix of Jordan diagonial form which contains the eigenvalues of matrix A.

As we understand the whole problem (the exact form of solution and the ultimate behaviour of the solution) is based on the eigenvalues of matrix A.

Considering the matrix A of our problem i.e.

$$A = \begin{bmatrix} 1 & -e\varepsilon \\ 1 & 1 \end{bmatrix}$$

we obtain the characteristic polynomial $\varphi(\lambda)$.

$$\varphi(\lambda) = \det(\lambda I - A) = \left| \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 1 & -e\varepsilon \\ 1 & 0 \end{bmatrix} \right| = \begin{vmatrix} \lambda - 1 & e\varepsilon \\ -1 & \lambda \end{vmatrix} = 0$$

Hence, $\lambda(\lambda - 1) + \varepsilon e = 0 \leftrightarrow \lambda^2 - \lambda + \varepsilon e = 0$ (2.9.1)

The last equation (which determines the eigenvalues of A and the solution of the system) is exactly the same with the equation which Benjamin (1984) discuss in his paper. The discussion about the roots of the equation will become parallel with the paper mentioned above.

Finally we may calculate the eigenvectors in order to find the Q matrix (and consequently Q^{-1}). Obviously the eigenvectors and consequently matrix Q has no influence to the ultimate behaviour of the solution.

The approach discussed above may be characterized as the direct approach for calculating A^n. There is also another one (indirect) which is based on the z-tranformation (see Appendix I). As we can easily identify

$$\Phi(n+1) = A \cdot \Phi(n) \text{ and } \Phi(0) = I$$

If we take the z-transformation of the first relationship we obtain,

$$\mathcal{Z}\{\Phi(n+1)\} = \mathcal{Z}\{A \cdot \Phi(n)\} \Leftrightarrow$$

$$z \cdot \mathcal{Z}\{\Phi(n)\} - z\Phi(0) = A \cdot \mathcal{Z}\{\Phi(n)\} \Leftrightarrow$$

$$[zI - A] \cdot \mathcal{Z}\{\Phi(n)\} = z\Phi(o) = zI \Leftrightarrow$$

$$\mathcal{Z}\{\Phi(n)\} = z[zI - A]^{-1} \Leftrightarrow$$

$$\Phi(n) = \mathcal{Z}^{-1}\{ z \cdot [zI - A]^{-1}\}$$

($\mathcal{Z}$ notation is explained in Appendix I).

Gadjow (1973) proposed a very useful algorithm in order to calculate $[zI-A]^{-1}$ and consequently $\Phi(n)$.

2.10 Transfer matrix (function) of the system

Another approach to the description and solution of dynamic systems is the respective transfer function. The transfer function relates directly, the z (or Laplace) transformed input variables with the z (or Laplace) transformed output variables.

If we apply the z-transformation on the equations of the standard system we obtain

$$\left.\begin{matrix} \mathcal{Z}\{\underline{x}_{n+1}\} = \mathcal{Z}\{A \cdot \underline{x}_n + B \cdot \underline{u}_n\} \\ \mathcal{Z}\{\underline{y}_n\} = \mathcal{Z}\{C \cdot \underline{x}_n + D \cdot \underline{u}_n\} \end{matrix}\right\} \rightarrow \left.\begin{matrix} z[\underline{x}_z - \underline{x}_0] = A \cdot \underline{x}_z + B\underline{u}_z \\ \underline{y}_z = C \cdot \underline{x}_z + D\underline{u}_z \end{matrix}\right\}$$

(Assuming (a) zero initial i.e. $\underline{x}_0 = \underline{0}$ and

(b) $\underline{x}_z = \mathcal{Z}\{\underline{x}_n\}, \underline{y}_z = \mathcal{Z}\{\underline{y}_n\}, \underline{u}_z = \mathcal{Z}\{\underline{u}_n\}$

Hence,

$$\left.\begin{matrix} z \cdot \underline{x}_z = A \cdot \underline{x}_z + B\underline{u}_z \\ \underline{y}_z = C \cdot \underline{x}_z + D\underline{u}_z \end{matrix}\right\} \rightarrow \left.\begin{matrix} (zI - A)\underline{x}_z = B \cdot \underline{u}_z \\ \underline{y}_z = C \cdot \underline{x}_z + D\underline{u}_z \end{matrix}\right\} \rightarrow$$

$$\rightarrow \left.\begin{matrix} \underline{x}_z = (zI - A)^{-1} \cdot B \cdot \underline{u}_z \\ \underline{y}_z = C \cdot (zI - A)^{-1} \cdot B\underline{u}_z + D\underline{u}_z \end{matrix}\right\} \rightarrow \left.\begin{matrix} \underline{x}_z = (zI - A)^{-1} \cdot B \cdot \underline{u}_z \\ \underline{y}_z = [C \cdot (zI - A)^{-1} \cdot B + D] \cdot \underline{u}_z \end{matrix}\right\}$$

The polynomial matrix

$$G(z)=C[zI-A]^{-1}B+D \tag{2.10.1}$$

is called the transfer matrix function of the system. (The discussion above is based on Gadjow (1973). Working similarly as above we may obtain the transfer function of our system which has the same form as in equation (2.10.1) (although our system is slightly different from the standard format).

Using the matrices A,B,C,D of our example we shall calculate the transfer matrix function of our system.

Firstly, we shall calculate the $[zI-A]^{-1}$, i.e.

$$[zI-A]^{-1}=\begin{bmatrix} z-1 & e\varepsilon \\ -1 & z \end{bmatrix}^{-1}=\frac{1}{z^2-z+\varepsilon e}\cdot\begin{bmatrix} z & -e\varepsilon \\ 1 & z-1 \end{bmatrix}$$

(Generally, if $Q=\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$ then $Q^{-1}=\dfrac{1}{\det(Q)}\begin{bmatrix} \delta & -\beta \\ -\gamma & \alpha \end{bmatrix}$)

consequently we obtain G(z) from equation (2.10.1) i.e.

$$G(z)=[0 \quad -\varepsilon]\cdot\frac{1}{z^2-z+e\varepsilon}\begin{bmatrix} z & -e\varepsilon \\ 1 & z-1 \end{bmatrix}\cdot\begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}+\begin{bmatrix} 0 & 0 & \frac{1}{e} \end{bmatrix}=\cdots$$

$$=\begin{bmatrix} \dfrac{\varepsilon}{z^2-z+e\varepsilon} & 0 & \dfrac{1}{e}-\dfrac{\varepsilon}{z^2-z+e\varepsilon} \end{bmatrix}$$

2.11 Input Signals - Output Responses

As we have seen in section (2.6) there are three types of variables (Input - state and output). Having developed the theory of transfer function, we may ignore at a first stage the state variable and examine the relationship between input and output through this new bridge (the transfer function).

Keeping also in mind that the transfer function theory has been developed by using z-transformation we may picture the following diagram.

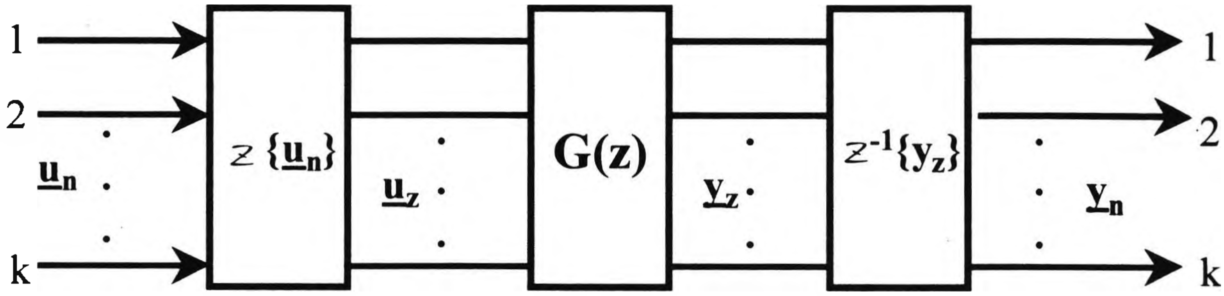


Diagram (2.11.1.)

We may immediately have a first feeling of the behaviour of the output caused by different types of input signals. That is obtained in two ways.

(a) Simulating numerical values for an input signal and calculating (following the last diagram 2.11.1) the output values. That is very useful to examine the behaviour of the system over a finite time period.

(b) Using the analytical form of the equation of the transfer function $\underline{y}_z = G(z) \cdot \underline{u}_z$ (after having obtain the z-transformation for input variable) and determine the exact pattern of the output values up to infinity. Of course this method is better and give us the safest answer for the ultimate behavior of the system.

Control theorists use standard input signals to run a first test for a dynamic system. Some of those are given in the table below (including z-transformation). (see Appendix I).

Name	Pattern $(x_0,x_1,x_2,...)$	z-transformation
spike	1,0,0,0,...	1
step	1,1,1,1,...	$\frac{1}{1-z^{-1}}$
ramp	0,1,2,3,...	$\frac{z^{-1}}{(1-z^{-1})^2}$
sine	$\sin\varphi$, $\sin(w+\varphi)$,...	$\frac{(\sin\omega)z^{-1}}{1-2(\cos\omega)z^{-1}+z^{-2}}$
geometric	$1,a,a^2,a^3,...$	$\frac{1}{1-az^{-1}}$
delay	x_{t-m}	$z^{-m}x_z$

All the above signal patterns may have a verbal interpretation which is applicable to our problem. For example considering the spike signal we may translate it as follows: "How the system reacts if we have an unexpected claim of one unit".

The step signal may be interpreted as follows:

The underwriter has made a mistake when he assessed the risk for the policy so the actual claims will be one unit greater for any future time value t. How the system will react?

The ramp signal or the geometric one may be used to model an increasing trend in mortality or morbidity rates of a group of insured lives or claim frequency for a group of policies. So the application of such a signal as an input variable will reveal the future behavior and the possible inertia mechanism of the premium rating procedure of our system. If the results are not in line with our expectations we may change the premium rating in order to avoid undesirable future results.

Another question which also arises is the following: Has the system been designed properly to obtain a final stable state? All the questions above should be answered using the techniques (a) and (b) described earlier in this section.

2.12 Non Linear problems & linearization procedure

Our discussion up to now, has considered only linear problems and consequently linear systems of equations. Of course there is something more, the non-linear systems.

Generally speaking, nonlinearities are divided into inherent and intentional ones. Ogata (1970) provides a certain list of the inherent non-linearities (e.g. Saturation, Dead zone etc.).

Intentional nonlinearities are introduced in order to obtain sophisticated models with high performance attitude.

Now, although there is a great need to establish analogous theory for non-linear systems, mathematicians have been unwilling to do so because of two reasons:

(a) Non-Linear theory appears extremely difficult and sometimes impossible to be handled.

(b) It has been established a standard linearization procedure by which a non-linear system may be easily converted to a linear one and be studied using standard linear techniques providing enough insight and accuracy to our questions.

Recently, there is a lot of discussion and some objections have been raised against point (a) as there are many scientists (in different areas) who have been involved with a newly developed theory (known by the Greek word "chaos") which studies non-linear systems.

We stop here the discussion for chaos (perhaps that may form another Ph.D. thesis... chaos in insurance!!) and develop a standard technique to linearize a non-linear system. We shall develop the linearizaton procedure for the continuous type of a dynamic system (the work for the discrete type is exactly parallel, using differences instead of derivatives). The non-linear vector form of the equation (2.6.5) is given below:

$$\dot{\underline{x}}(t)=\underline{f}(\underline{x}(t),\underline{u}(t)),\quad \underline{x}(t)\in\mathbb{R}^{h},\quad u(t)\in\mathbb{R}^{m} \tag{2.12.1}$$

The linearization of a system is usually obtained over the equilibrium points $\underline{x}*$ (which is related with constant input signals $\underline{u}*$=constant).

Considering equation (2.12.1) we may define the equilibrium points $\underline{x}*$ (for the $\underline{u}*$ constant input) as the solution of equation below:

$$\underline{f}(\underline{x}*,\underline{u}*)=\underline{O} \tag{2.12.2}$$

Then we consider small deviations i.e.

$$\underline{x}=\underline{x}*+\delta\underline{x} \quad \text{and} \quad \underline{u}=\underline{u}*+\delta\underline{u}$$

and we expand a Taylor series over the point $(\underline{x}*,\underline{u}*)$ for the function $\underline{f}(\underline{x}(t),\underline{u}(t))$ obtaining.

$$\underline{f}(\underline{x},\underline{u})=\underline{f}(\underline{x}*,\underline{u}*)+\left.\frac{\partial\underline{f}}{\partial\underline{x}}\right|_{(\underline{x}*,\underline{u}*)}\cdot\delta\underline{x}+\left.\frac{\partial\underline{f}}{\partial\underline{u}}\right|_{(\underline{x}*,\underline{u}*)}\cdot\delta\underline{u}+\ldots \tag{2.12.3}$$

where

$$\left.\frac{\partial\underline{f}}{\partial\underline{x}}\right|_{(\underline{x}*,\underline{u}*)}=\left.\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_h}{\partial x_1}\\ \vdots & & \vdots\\ \frac{\partial f_1}{\partial x_h} & \cdots & \frac{\partial f_h}{\partial x_h}\end{bmatrix}\right|_{(\underline{x}*,\underline{u}*)}=A\in\mathbb{R}^{h\times h}$$

$$\frac{\partial \underline{f}}{\partial \underline{u}}\Big|_{(\underline{x}^*,\underline{u}^*)} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_h}{\partial u_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial u_m} & \cdots & \frac{\partial f_h}{\partial u_m} \end{bmatrix}\Bigg|_{(\underline{x}^*,\underline{u}^*)} = B \in \mathbb{R}^{h\times m}$$

Ignoring the terms (with power ≥ 2) of the Taylor expansion in equation (2.12.3) we obtain the final form of the system.

$$\delta \underline{\dot{x}} = A\delta \underline{x} + B\delta \underline{u} \tag{2.12.4}$$

which describes the linear approximation of $\underline{\dot{x}} = f(\underline{x},\underline{u})$ over the area of the equilibrium point $(\underline{x}^*,\underline{u}^*)$. The discussion above is based on Kalogeropoulos (1987), Ogata (1970) and Gagzow (1973).

2.13 Controllability (Definitions - Theorems - Applications)

In this section, we shall discuss a very important concept of the linear dynamic system. It is a qualitative property, named controllability which has been introduced by Kalman and plays an important role (along with the other concept of observability) in the optimal control. Most of the physical systems are controllable, but the corresponding mathematical models may not possess the property. Here we shall develop the necessary conditions for questioning the certain property.

Generally speaking, there are two types of controllability: the state and the output one. We shall first discuss the state controllability providing the necessary definitions and theorems.

Definition (2.13.1) (Athans & Falb (1966)): A state $\underline{x}_1$ is said to be reachable, or accessible, from the state $\underline{x}_0$ at t_0 with respect to U if there is an element u_t ($\underline{u}_t$, $t \in [t_0, t_1]$) of U such that

$$\underline{x}_{t_0} = \underline{x}_0 \quad \text{and} \quad \underline{x}_{t_1} = \underline{x}_1$$

for some finite $t_1 \geq t_0$

where U is the set of the admissible controls.

Definition (2.13.2) (Athans & Falb (1966)): If the state $\underline{x}_1 = \underline{0}$ is reachable from $\underline{x}_0$ at t_0, then we say that $\underline{x}_0$ is controllable at (time) t_0. In other words $\underline{x}_0$ is controllable at t_0 if there is a piecewise continuous function u^0 defined over $[t_0,T]$ such that

$$\underline{x}_T = \underline{0}$$

Definition (2.13.3) (Athans & Falb (1966)): If every state x_0 is controllable at time t_0 then we say that the system is controllable at t_0. If every state $\underline{x}_0$ is controllable at every time t_0 in the interval of definition of the system then we say that the system is completely (state) controllable.

We shall also quote the verbal interpetation of state controllability of Athans & Falb (1966).

"Controllability means that it is possible to drive any state of the system to the origin in some finite time".

A theorem will be given below which operates as a criterion in order to check the concept of complete state controllability of the system, in the linear time invariant system (discrete or continuous type).

Theorem (2.13.1) (Athans & Falb (1966)). Let the general form of a dynamic linear system S(A,B,C,D) then the system (A,B) is complete controllable if and only if Rank $\left[\, B \vdots AB \vdots A^2B \vdots \ldots \vdots A^{h-1}B \,\right] = h$ where h is the number of rows (columns) of matrix A.

The matrix $\mathscr{B} = \left[\, B \vdots AB \vdots \ldots \vdots A^{h-1}B \,\right]$ is called matrix of controllability.

If a system is not complete controllable then it is proved that the maximum controllable space is given by X_c=col·span $[\,B\vdots AB\vdots\ldots\vdots A^{h-1}B\,] \subseteq \mathbb{R}^h$ (col-span means the vector supspace determined by the columns of the specific matrix). Another useful theorem is the following.

Theorem (2.13.2) (Kalogeropoulos (1987)). The controllability property is not destroyed by linear transformations.

The above theorem is very useful in conjunction with the concept of feedback action. As we are going to see in the next sections feedback (i.e. a linear transformation) will be used to redesign a system and achieve the required stability. Given the last theorem we may feel safe that the modification of (A,B) will not destroy controllability.

Another criterion (see Ogata (1970)) for checking the controllability property is based on the prospective diagonial form of matrix A. Let us assume the set of eigenvalues $\sigma(A)=\{\lambda_1,\lambda_2,\ldots,\lambda_n\}$ and $\lambda_i \neq \lambda_j$ $\forall i \neq j$ and $Q=[\underline{u}_1 \underline{u}_2 \cdots \underline{u}_n]$ the matrix of the respective eigenvectors then

(A,B) is complete controllable $\Leftrightarrow$ All the rows of $Q^{-1}\cdot B$ are different from zero.

Now, we shall also provide a necessary and sufficient condition for complete state controllability in the z-plane in terms of the transfer function (Ogata (1970)).

A system is completely state controllable if and only if "no cancellation occurs in the transfer function (or transfer matrix). If cancellation occurs the system cannot be controlled in the direction of the cancelled mode".

Finally we shall state the condition for the completely output controllability (Ogata (1970)).

The system $S_d(A,B,C,D)$ (or $S_c(A,B,C,D)$) is completely output controllable if and only if the $m \times \ell$ matrix

$$\left[\ CB \vdots CAB \vdots CA^2B \vdots \ldots \vdots CA^{h-1}B \vdots D\ \right]$$

is of rank m.

Examination of state controllability of our model

Considering our model, we may apply the basic theorem and examine the controllability of the system

$$A=\begin{bmatrix}1 & -e\varepsilon \\ 1 & 0\end{bmatrix}, B=\begin{bmatrix}-1 & 0 & 1 \\ 0 & 0 & 0\end{bmatrix}, A\cdot B=\begin{bmatrix}-1 & 0 & 1 \\ -1 & 0 & 1\end{bmatrix}$$

$$\text{rank } [B \vdots AB]=\text{rank}\begin{bmatrix}-1 & 0 & 1 & \vdots & -1 & 0 & 1 \\ 0 & 0 & 0 & \vdots & -1 & 0 & 1\end{bmatrix}=2$$

Hence, our system is complete state controllable.

Examination of output controllability of our model

$$C=\begin{bmatrix}0 & -\varepsilon\end{bmatrix}\ ,\ D=\begin{bmatrix}0 & 0 & \frac{1}{e}\end{bmatrix}$$

$$C\cdot B=\begin{bmatrix}0 & 0 & 0\end{bmatrix}\ ,\ CAB=\begin{bmatrix}\varepsilon & 0 & -\varepsilon\end{bmatrix}$$

$$\text{rank}[CB \vdots CAB \vdots D]=\begin{bmatrix}0 & 0 & 0 & \vdots & \varepsilon & 0 & -\varepsilon & \vdots & 0 & 0 & \frac{1}{e}\end{bmatrix}=1$$

Hence the system is completely output controllable.

Comments: That means, we can fully control the system or guide the system from any initial state towards a desirable state in some finite time. (Normally we want to guide the system to $\underline{O}$).

Trying to obtain a further insight over the controllability concept we may say that:

- Controllability examines the behaviour of the system over a finite time interval.

- Stability (will be defined later in section (2.15)) examines the behavior of the system over the infinity.

Our basic aim should be to create a system with initial conditions which are included in the controllable subspace $X_c = \text{col}-\text{span}[\ B \vdots AB \vdots \ldots A^{h-1}B\ ]$ and consequently guide the system to the desired state. So avoid to put the initial conditions into the uncontrollable subspace.

2.14 Observability (Definitions - Theorems - Applications)

The concept of controllability refers to the first equation of the standard format of a dynamic system $S_d(A,B,C,D)$ as opposed to the concept of observability which refers to the second equation of system i.e.

$$\underline{y}_n = C \cdot \underline{x}_n + D \cdot \underline{u}_n\,,\ C \in \mathbb{R}^{m\times n}\,,\ D \in \mathbb{R}^{m\times \ell}$$

Definition (2.14.1) (Athans & Falb (1966)): We say that a state $\underline{x}_0$ is observable at t_0 if, given any control u, there is time $t_1 \geq t_0$ such that knowledge of $\underline{u}_{[t_0,t_1]}$ and the output $\underline{y}_{[t_0,t_1]}$ is sufficient to determine $\underline{x}_0$.

Definition (2.14.2) (Athans & Falb (1966)): If every state $\underline{x}_0$ is observable at t_0 then we say that the system is observable at t_0. If every state $\underline{x}_0$ is observable at every time t_0 in the interval of definition of the system, then we say that the system is completely observable.

We shall quote again the verbal interpretation of Athans & Falb (1966) for observability i.e. "observability means that the initial state of the system can be found from a suitable measurement of the output".

An analogous theorem as in the previous section will be described below which operates as a criterion in order to check the concept of complete observability of the system.

Theorem (2.14.1) (Athans & Falb (1966)) The system S(A,B,C,D) is complete observable. If and only if

$$\operatorname{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{h-1} \end{bmatrix} = h$$

The matrix $\mathscr{C} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{h-1} \end{bmatrix}$ is called matrix of observability.

If a system is not complete observable then it may be proved that the maximum observable space is given

$$\hat{\mathscr{C}} = \operatorname{Nr}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{h-1} \end{bmatrix} \underline{\subset} \mathbb{R}^h$$

(where $\operatorname{Nr}(F) = \{\underline{x} \in \mathbb{R}^h : F(\underline{x}) = \underline{O}\}$).

Another parallel theorem for observability is the following:

Theorem (2.14.2) (Kalogeropoulos (1987)). The observability property is <u>not</u> destroyed by linear transformations.

So, the linear transformation of a certain feedback action will not destroy observability of the system.

The last criterion applied to prospective diagonial form of the matrix A is the following:

Theorem (2.14.3) (Ogata (1970)). The dynamic system S(A,B,C,D) is complete controllable if and only if all the columns of C·Q are different from zero. (Q has been defined in section (2.13) for a controllability theorem).

Finally we should mention the conditions for complete observability in the z-plane which is exactly the same as for the complete controllability (i.e. No cancellation should occur in the transfer function or transfer matrix).

Examination of observability of our model

Considering our problem we obtain

$$A = \begin{bmatrix} 1 & -e\varepsilon \\ 1 & 0 \end{bmatrix}, \; C = \begin{bmatrix} 0 & -\varepsilon \end{bmatrix}, \; CA = \begin{bmatrix} -\varepsilon & 0 \end{bmatrix}$$

Hence rank $\begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 0 & -\varepsilon \\ -\varepsilon & 0 \end{bmatrix} = 2$

that means our system is complete observable.

Comments:

As we have stressed, observability is a very important property of a linear dynamic system. But what is the "meaning" of it, what is the physical interpretation?

Let us think the meaning of word "observability" in general... "ability" to observe" and the three spaces input-state-output. Obviously we can observe the first and the last one. It is not obvious that we can observe the state space. This space is generally hidden and describes the inertia mechanism. So it is possible to obtain systems totally or partially unobservable. Of course this is not a desirable situation, and we are trying to avoid those systems. If we have to work with those systems we shall target to stay within the observable part of the state space.

Finally, we shall quote again Athans & Falb (1966) in order to explain the need of observability property. "Suppose that the system is observable this implies that we can compute the initial state x_0 at t=0". So we can fully regulate the output maintaining it at zero (after guiding it to zero, assuming the controllability property for a dynamic system).

2.15 Stability (Definitions - Theorems - Applications)

As we have already pointed out in the comments of section (2.13) stability is another important property of linear dynamic systems which examines the behaviour of the system "near the infinity area". Diagrammatically

$$\boxed{\text{controllability concept}} \xrightarrow[n\to\propto]{} \boxed{\text{stability concept}}$$

Of course the basic target is the design of stable systems. That means we have to establish certain mechanisms into the system in order to obtain bounded output responses whatever is the input signals. (A special case of the previous concept is the B.I.B.O. system i.e. Bounded Input Bounded Output).

Stability is usually examined over the points which are called equilibrium points (zero $\underline{0}$ can always be one of them, perhaps after a suitable linear transformation). The proof is obvious considering the following definition of an equilibrium point.

Definition (2.15.1) (Ogata (1970)): A point $\underline{x}_e \in X$ is called equilibrium point (state) of the system if and only if by definition $\underline{x}_t = \underline{x}_e \ \forall t \geq 0$ given that $\underline{x}_0 = \underline{x}_e$ and $\underline{u}_t = \underline{0} \ \forall t \geq 0$ i.e. $\underline{x}_e$ is an equlibrium state if the solution of the system $\underline{x}_t$ remains on it as far as we have zero input signal.

Now considering the equation $\underline{x}_{n+1} = A \cdot \underline{x}_n + B \cdot \underline{u}_n$ we obtain

$$\underline{x}_e \text{ equilibrium point} \Leftrightarrow A \cdot \underline{x}_e = \underline{x}_e .$$

$\underline{0}$ may always be a solution of the last equation. Now if A is a non-singular matrix $\underline{0}$ is the only equilibrium point (if A is singular there are infinite solutions).

There are two definitions for stable points: i.e.

Definition (2.15.2): An equilibrium point (state) $\underline{x}_e$ is called (Liapunov) stable point if and only if by definition

$$\forall S(\underline{x}_e,\xi) = \{\underline{x} \in X : \|\underline{x} - \underline{x}_e\| < \xi, \xi > 0\} \quad \exists \delta > 0 : S(\underline{x}_e,\delta) \subset S(\underline{x}_e,\xi)$$

Such that for every $\underline{x}_0 \in S(\underline{x}_e,\delta)$ then $\underline{x}_t \in S(\underline{x}_e,\xi) \quad \forall t > t_0$

that means the state vector remains always "near" the equlibrium point or quoting from Ogata (1970), "corresponding to each $s(\xi) = s(\underline{x}_e,\xi)$, there is an $s(\delta) = s(\underline{x}_e,\delta)$ such that trajetories starting in $s(\delta)$ do not leave $s(\xi)$ as t increases indefinitely. The real number δ depends on ε and in general also depends on t_0. If δ does not depend on t_0 the equilibrium state is said to be uniformly stable".

Definition (2.15.3): An equilibrium state $\underline{x}_e$ is called stable asymptotic point if and only if by definition

$\underline{x}_e$ is a Liapunov stable point (according to definition (2.15.2))

and $\lim\limits_{n \to \infty} \underline{x}_n = \underline{x}_e$ (or $\lim\limits_{n \to \infty} \|\underline{x}_n - \underline{x}_e\| = 0$).

that means the state vector remains always "near" the equilibrium point while converging to it. Theoretically the state vector equals the equilibrium point near the infinity area.

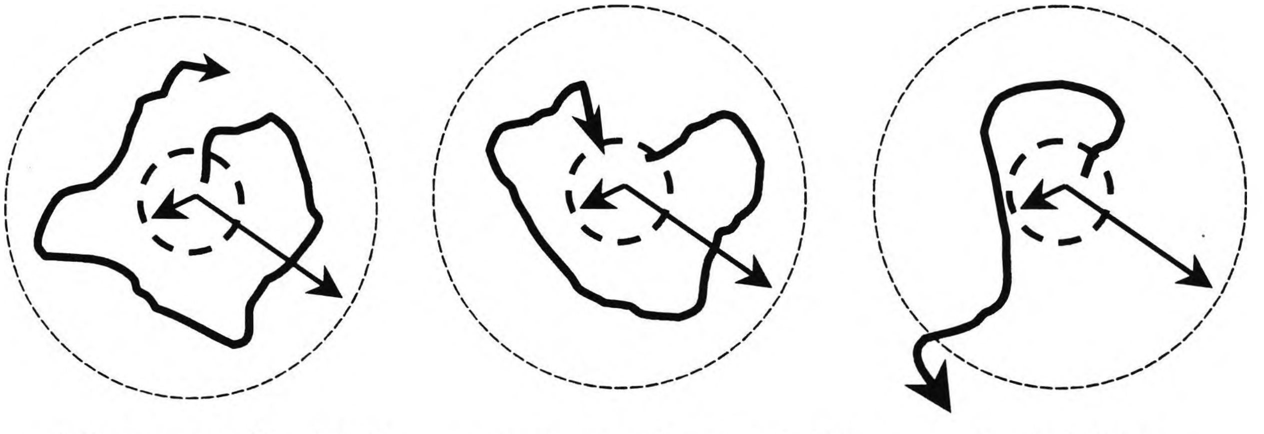
The asymptotic stability is a local concept, so most times we need to find the largest region where this property holds for the system. This region is called the domain of attraction. Actually it is the part of the state space where every trajectory originated is asymptotically stable.

We shall also provide the formal definition of instability (Ogata (1970)).

Definition (2.15.4): An equilibrium state $\underline{x}_e$ is said to be unstable if for some real number $\xi>0$ and any real number $\delta>0$, no matter how small, there is always a state $\underline{x}_0$ in $S(\delta)$ such that the trajectory starting at this state leaves $s(\xi)$.

That means the state vector does not remain "near" $\underline{x}_e$.

The following diagram provides a geometrical representation of the two definition compared with the unstable situation. The diagram (2.15.1) shows a typical path for each of the three situations.



Liapunov Stability **Asymptotic Stability** **Instability**

Diagram (2.15.1)

There is a basic theorem operating as a first criterion in order to judge the stability of a system. That involves the eigenvalues of matrix A. We shall give the two versions for discrete and continuous type of systems.

Theorem (2.15.1): (Stability Criterion) (Kalogeropoulos (1987) and Ogata (1970)). The equilibrium point $\underline{x}_e$ is said to be:

(a) Liapunov stable point if and only if all the eigenvalues of matrix A (where a+bi is the general form of an eigenvalue of matrix A).

 (i) (continuous) have non-positive real part (i.e. $a \leq 0$) and those with zero real part (i.e. a=0) have simple structure (see Appendix II).

 (ii) (discrete) have absolute value less or equal to unity (i.e. $|a+bi| \leq 1$) and if $|a+bi| = 1$ then the a+bi should have simple structure.

(b) Asymptotic stable point if and only if all the eigenvalues of matrix A (with the complex form a+bi)

 (i) (continuous) has negative real part (i.e. $a<0$)

 (ii) (discrete) has absolute value less than unity (i.e. $|a+bi| < 1$).

(c) Unstable point if and only if there is at least one eigenvalue of matrix A.

 (i) (continuous) with positive real part, or with zero real part and not simple structure.

 (ii) (discrete) with absolute value greater than unity or equal to unit and not simple structure.

Obviously the stability of the system depends upon the magnitude of the eigenvalues of matrix A. Consequently the eigenvalues depend upon the roots of the characteristic polynomial of A i.e. the roots of $\varphi(\lambda)=\det(\lambda I-A)=0$.

The $\varphi(\lambda)$ is an h^{th} order polynomial and (generally) it is very difficult to be solved analytically. Control theorists have developed, a very powerful tool called root locus method in order to examine the position of the roots in the z-plane. Root locus method and other criteria will be discussed in Appendix III.

Example and comments

Considering our example we obtain,

$$\varphi(\lambda)=\det(\lambda I-A)=\begin{vmatrix}\lambda-1 & e\varepsilon \\ -1 & \lambda\end{vmatrix}=\lambda^2-\lambda+\varepsilon e=0 \qquad (2.15.1)$$

We have found again this equation above in section (2.9) in the calculation of A^n.

We generally know that $\rho_1 \cdot \rho_2 = \varepsilon e$ (ρ_1,ρ_2 roots of equation (2.15.1)). So if $|\varepsilon e|>1$ then $|\rho_1\rho_2|>1$ and consequently there exists at least one root ρ_1 or ρ_2 with absolute values greater than unity and the system is unstable. Hence, a first action would be to design the system such that $|\varepsilon e|<1$.

Of course we must examine the equation more accurately and answer other questions. Which is the best selection for the parameters? Is there a best one? How we determine the concept of "best". All these questions will be answered in the next two sections in the context of the system feedback action and best selection of parameters according to certain criteria.

2.16 Feedback Action - Redesign of poles to obtain the required stability of the system

Another basic concept of linear dynamic systems is the "feedback action". It is directly related with the state space and is usually represented by a linear transformation F:X→U where $\underline{u}(t)=F\cdot\underline{x}(t)+\underline{v}(t)$ and $\underline{v}(t)$ is a new input function. So the initial system S(A,B) is transformed to the S(A+BF,B).

But what is the need of feedback action? The theorem below provides a first answer.

Theorem (2.16.1) (Kalogeropoulos (1987)): Let the dynamic system S(A,B,C,D) and F:X→U, linear transformation.

(i) col-span $[\ B\vdots AB\vdots\ldots\vdots A^{h-1}B\ ]$=col-span $[\ B\vdots(A+BF)B\vdots\ldots\vdots(A+BF)^{h-1}B\ ]$ i.e. the controllability property is not destroyed.

(ii) $$\mathrm{Nr}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{h-1} \end{bmatrix} = \mathrm{Nr}\begin{bmatrix} C \\ C(A+BF) \\ \vdots \\ C(A+BF)^{h-1} \end{bmatrix}$$

i.e. the observability property is not destroyed.

(iii) If the system S(A,B,C,D) is complete controllable then the set of eigenvalues of (A+BF) or the poles of S(A+BF,B,C,D) may be arbitrarily designed by choosing a suitable F.

Hence if a system is controllable but not stable we may design the poles in order to achieve the desired stability (remaining controllable). The poles refer to the roots of the characteristic polynomial of the matrix A or also may refer to the roots which make zero the denominator of the transfer function in a simple input - simple output dynamic system.

Apart from the state feedback action there is the output feedback action where the output information is measured and used to stabilize the system.

Let L be an output feedback action,

$$L : Y \rightarrow U \ : \underline{u}(t) = L\underline{y}(t) + \underline{v}(t)$$

It is easily proved that the output feedback action may be translated into a state feedback action.

There are a lot of types of feedback mechanisms depending on how they use the historical information of the system (history of the state vector). We shall discuss three of the basic ones.

1. Proportional Action:

Considers the difference between the value of the state vector at the present time and the respective desired value of the state and feeds back a proportion to the system. In our example, we have a proportional feedback action considering only the final value of the accumulated surplus as the respective desired value is zero.

2. Integral Action:

Considers the whole history of the state vector (generally is based on integral (continuous type) or summation (discrete type) of the state vector).

3. Derivative Action:

Considers the short history immediately before the present time. Generally speaking is based on the derivative or difference of state vector variable.

Of course, the three actions may be combined and produce a synthesis which will result better performance for the system, so we have

- proportional - plus - integral controllers
- proportional - plus - derivative controllers
- proportional - plus - derivative - plus - integral controllers.

Application to our example

In our problem there is a feedback mechanism and it is actually a proportional one. The choice of matrix F (see notation before) is restricted to the choice of ε i.e. the certain proportion of accumulated surplus which should be fed back to the system. So if we want a stable system we should choose ε such that the poles (or the eigenvalues of matrix A) are within the unit circle.

In the next section we shall examine another use of the feedback action with respect to optimization problems.

2.17 Optimal and adaptive Control in dynamic systems

The optimal control theory is very extensive and in nowadays plays the most important role as there is a great demand for systems which perform in the best possible way. In Appendix IV, there is a short discussion of the required knowledge from standard analysis and functional analysis with respect to optimization techniques.

In this section we shall discuss the linear - time - invariant optimal control systems which are based on quadratic performance indexes along with the concept of feedback.

Actually we shall quote the result from Ogata (1970).

Let us consider a certain dynamic system

$$\dot{\underline{x}} = A \cdot \underline{x} + B \cdot \underline{u} \tag{2.17.1}$$

and the need for optimization of the quadratic performance index

$$J(u) = \int_0^{\infty} \left(\underline{x}'(t) Q \underline{x}(t) + \underline{u}'(t) R \underline{u}(t) \right) dt \tag{2.17.2}$$

where Q is a positive - definite (see Appendix IV for the definition) (or positive - semidefinite) real symmetric matrix and R is a positive - definite real symmetric matrix. Then it is proved (Ogata (1970)) that

$$\underline{u}(t) = -K \underline{x}(t) \tag{2.17.3}$$

is the optimal control law and K matrix is determined by the following equation

$$K = R^{-1} \cdot B' \cdot P \tag{2.17.4}$$

P matrix is the solution in the reduced - matrix Riccati equation (see equation (2.17.5)).

$$A'P + PA - PBR^{-1}B'P + Q = 0 \tag{2.17.5}$$

(The (′) superscript stands for transpose matrix).

Hence, feedback action may produce an optimal system (apart from stable which we have seen in the previous section).

Before closing this section we shall briefly refer to adaptive control systems, quoting from Ogata (1970).

"Adaptation is a fundamental characteristic of living organisms since they attempt to maintain physiological equilibrium in the midst of changing environmental conditions. An approach to the design of adaptive systems in then to consider the adaptive aspects of human or animal behavior and to develop systems which behave somewhat analogously".

Obviously, from the definition above the adaptive systems are the most interesting ones. Actually, we always target the designation of such a system.

So, is it possible to produce the ideal adaptive systems? The answer is: No, because we can not ideally simulate the human's behavior. A manager may use other criteria to decide the next year's premium where sometimes are not clear even to himself!

2.18 Special Topics - Geometrical Representation

In this section, we shall try to concentrate in the geometrical representation of control theory while discussing some special topics.

a) Control Design and Diagrams

Having obtained the right model, we have solved the problem half way through. That means we should spend a lot of time and care modelling our problem. i.e. do the following:

(i) Determine the input variables.

(ii) Determine the output variables.

(iii) Determine the state variables (both the basic and subsidiaries ones) which should be followed and controlled in order to (finally) control the output.

(iv) Determine the mechanisms in the state space and the relationships between all the variables.

(v) Formulate a control diagram based on the established variables and relationships trying to obtain a further insight into the flow of the whole process.

(vi) Examine the basic properties controllability, observability and stability of the system.

(vii) Redesign (if necessary) the basic parameter values or create feedback mechanisms which may help for stability or optimality purposes.

(viii) Translate final results and make any required adjustments.

We provide below the diagram (2.18.1) of the general form of the system $S_c(A,B,C,D)$ (continuous type). Where A,B,C,D in the boxes are the respective matrices of equations (2.6.5) and (2.6.6) while the symbol in the central box stands for integration.

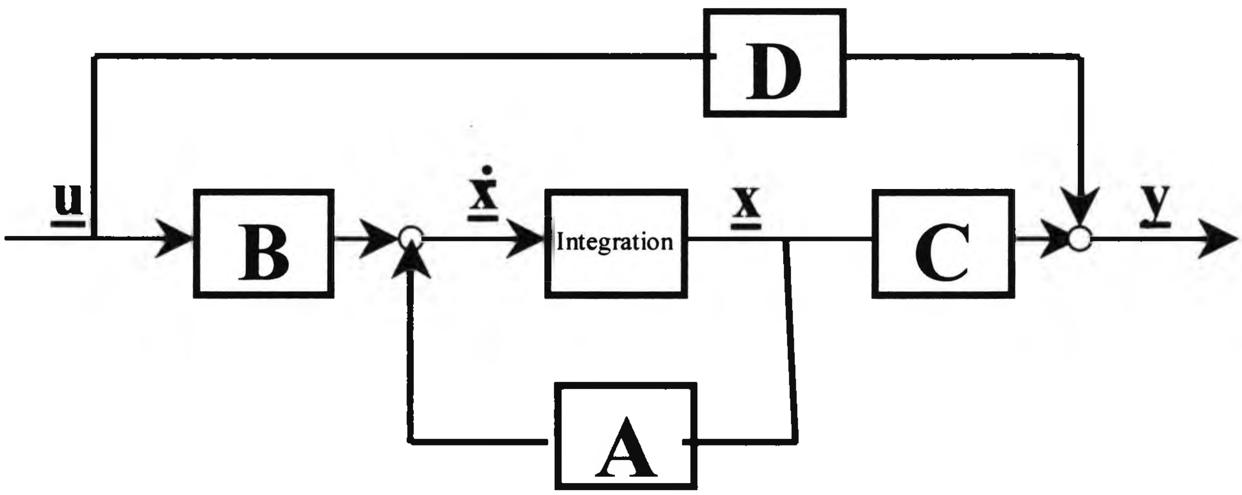


Diagram (2.18.1)

b) Transient Response Analysis

The transient response analysis of a system is very important and obviously predefines the ultimate state.

In order to examine the transient-response characteristics, we apply one or more of the standard signals and determine some special indices.

1. delay time, t_d
2. rise time, t_r
3. peak time, t_p

4. maximum overshoot, M_p
5. Settling time t_s

The interpretation of these indices are immediate just from the word description but we shall also quote the formal definitions from Ogata (1970).

Definition (2.18.1)

a. Delay time, t_d:... is the time required for the response to reach half the final value the very first time.
b. Rise time, t_r:... is the time required for the response to rise from 10 to 90%, of its final value.
c. Peak time, t_p:... is the time required for the response to reach the first peak of the overshoot (the response curve above the x-axis).
d. Maximum (percent) overshoot, M_p:... is the maximum peak value of the response curve measured from unity. If the final steady-state value of the response differs from unity then it is common to use the maximum percent overshoot. It is defined by the ratio

$$\frac{\text{Maximum Peak Value} - \text{Final Steady State Value}}{\text{Final State State Value}}$$

   The amount of M_p directly indicates the relative stability of the system.
e. Settling time, t_s:... is the time required for the response curve to reach and stay within a range about the final value of size specified by absolute percentage of the final value (usually 5% or 2%). The t_s is related to the largest time constant of the control system.

The definitions above are very important and useful in applications. Having established the concept of "settling time" we may answer the question raised in the previous section. "What is the best design of the system? Which is the best choice for a specific parameter? Obviously the smaller the settling time of a system the better design for our system.

c) Application to our example

Below we shall examine a simulation of our problem and present a geometric approach of the stability of the system drawing the path of the state vector in the state space (in our example the state space is the $\mathbb{R}^2$. Before we do so we shall determine the settling time for each scenario.

We consider a spike input signal and two different set of values for ε,e

(i) $\varepsilon=0.2$ $e=0.8$ (ii) $\varepsilon=0.5$ $e=0.8$

Then quoting the table from Benjamin (1984) we obtain the settling times for different scenarios as defined below.

Table (2.18.1)

n	(i): S_n	(ii): S_n
0	-1.000	-1.000
1	-1.000	-1.000
2	0.160	0.400
3	0.320	0.800
4	0.294	0.640
5	0.243	0.320
6	0.196	0.064
7	0.157	-0.064
8	0.125	-0.089
9	0.100	-0.064
10	0.080	-0.028
11	0.064	-0.002
12	0.051	0.008
13	0.041	0.009
14	0.032	0.006

We also determine as 5% the required percentage for the definition of settling time then.

The 5% settling time for the first scenario is something greater than 12 i.e.

$$t_{5\%}^{(i)} \cong 12 \tag{2.18.1}$$

while for the second scenario (using linear interpolation)

$$t_{5\%}^{(ii)} \cong 9.4 \tag{2.18.2}$$

So the second scenario is better than the first one considering the index of the settling time since

$$t_{5\%}^{(ii)} < t_{5\%}^{(i)}$$

but in the second scenario the system exhibits oscillations. So further research should be carried out in order to balance the minimum settling time with the minimum oscilation effects.

In the next diagram (2.18.1) we draw two paths one

(i) (blue path) "----" representation of $\underline{x}_n$ with the point (s_{n-1}, s_n) in the xy-plane

(ii) (red path) "— — — —" representation of $\underline{x}_n$ similarly as (i).

As we observe path (i) is closer to zero with no oscilation but goes slowly towards zero while path (ii) is not so close to zero, exhibits oscillations but goes faster towards to zero. After having obtained this graphical representation of the state of the system we may easily choose which is the best (according to our criteria) choice for the parameters involved and establish the respective strategy.

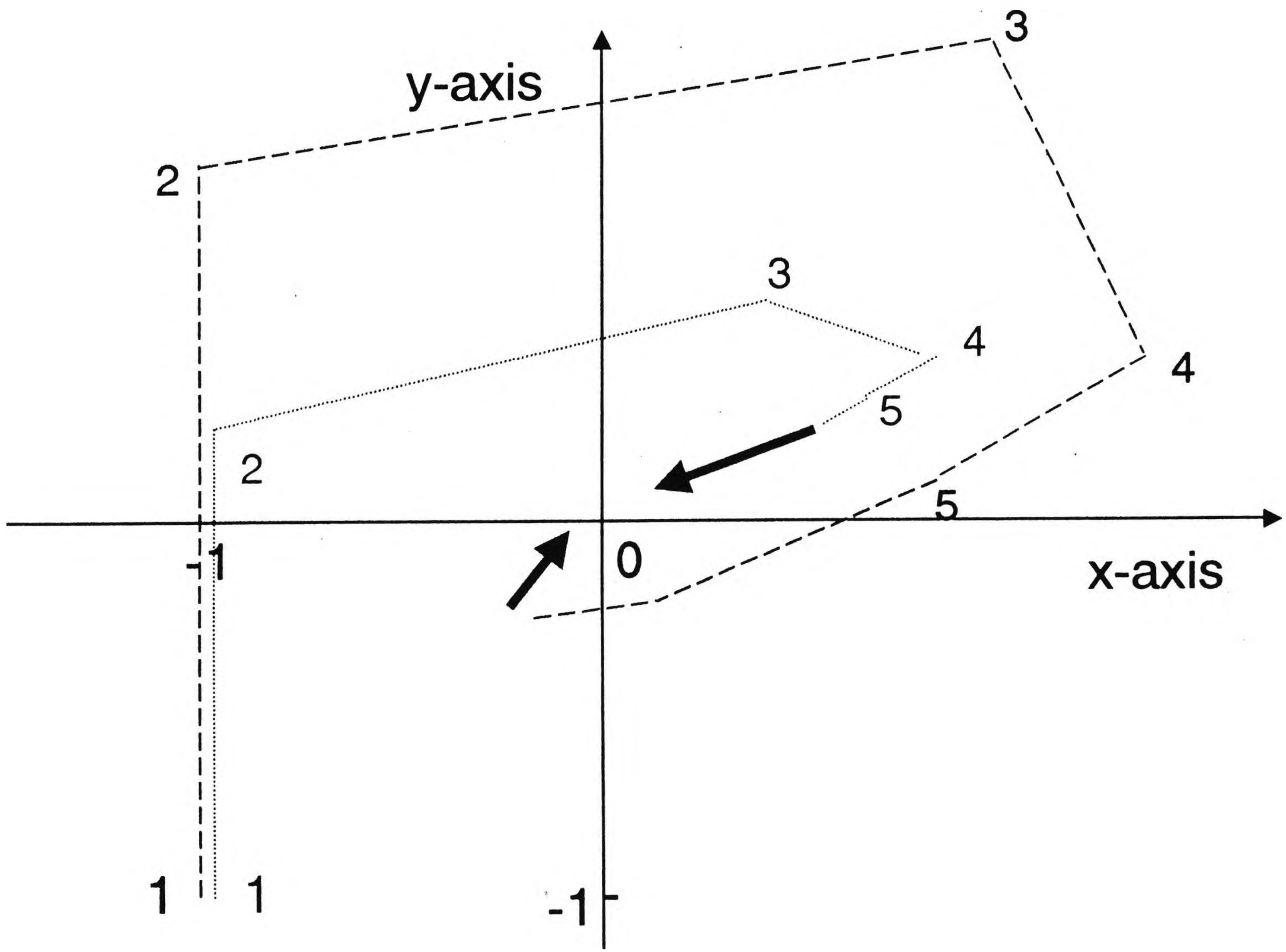


Diagram (2.18.2)

2.19 Final Remarks - Further Proposals

As we have already indicate in the introductory section of this chapter, our aim was not to provide a detailed analysis of control theory, but a summary of the most useful concepts and emphasize on the powerful tools of this theory which may be used by an actuary.

In this chapter we have presented the modern control theory approach using the vector form of the equations, the state space and linear algebra while providing also the basics from the conventional control theory approach with z-transformations and transfer function.

The modern approach is more powerful providing a very deep insight into the problem. The more theoretical form and the description of the system only with four matrices S(A,B,C,D) provide us unique flexibility and understanding of the inertia mechanism of the system.

Conventional control theory can handle efficiently single input - single output problems. Modern systems are much more complex. They are usually described as multiple input - multiple output or/and time-varying or/and non-linear ones.

The use of modern control techniques are unavoidable for the systems above.

Finally we should stress the superiority (supported also by Taylor (1987)) of the modern approach in the system design and especially in the optimal control problems with respect to given performance indices.

In the development of the thesis we shall use modern control theory while keeping some of the elementary approach of the conventional theory.

Chapter 3

Review of the Literature

(Applications of Control Theory to Insurance and Pensions)

3.1 Introduction

Having established the required theoretical background of control theory in chapter 2, we shall proceed now with the discussion of the relevant published paper work. It seems that many actuaries have pointed out the necessity for a new approach to the insurance problems as the traditional statistical methods or the individual / collective risk models appear to be not so practical or "safe".

Immediately after the development of Lundberg's ruin theory, De Finneti (1957) proposed a modified model for an insurance company using a simple premium control rule. As a continuation of this work Borch (1967) in a survey article for the history of risk theory identified the modern direction of risk theory into control techniques providing a general solution to De Finneti's model.

A paper of Seal (1970) presented in a Conference of the Society of Actuaries, provided the practical question and reasoning for developing the control techniques for insurance problems.

Afterwards and especially during the last two decades there are many persons who have been involved in modelling certain actuarial or insurance problems under the framework of control theory (starting from Balzer & Benjamin (1980)).

In the next sections we shall present (seperately, one by one) the paper work using the following structure.

(a) Description (of the paper, stating the general structure, the model used and the important results).

(b) Comments (concerning the results and the control techniques or the certain point of view of the author plus the links with previous research work. That will enable us to trace through the development of ideas).

The papers are presented in chronological order as we prefer to focus on the control concepts and how they have been developed and linked with the different areas (life insurance, general insurance, pensions) of the actuarial work.

Before closing the chapter, we shall provide a final section discussing the basic control concepts across all the papers in order to obtain an overview of the topic.

3.2 De Finetti (1957). "Su una impostazione alternativa della theoria collectiva del rischio"

(a) **Description**: As we have stated in the intoductory section, almost fifty years after the development of Lundberg's ruin theory, De Finetti has pointed out the unrealistic modelling structure of the collective risk model. The basic equation (3.2.1) which describes the surplus accumulation procedure drives the process up to infinity since

$$S_n = S_{n-1}+P_n-C_n \tag{3.2.1}$$

where S_n is the surplus at the end of year n

P_n is the annual charged premium in year n calculated using a safety loading θ_0 i.e.

$$P_n = (1+\theta_0)\cdot E(C_n) \tag{3.2.2}$$

C_n independent identically distributed random variables representing the total annual incurred claims in each year (n=1,2,...).

De Finetti proposed an upper limit for the surplus of the company (say Z). Actually his model is based on the assuptions below:

"(i) The company has an initial capital S_0.

(ii) In each operating period the company underwrites a portfolio of insurance contracts with a claim distribution $F(x)\cdot(F(x)=0$ for $x \leq 0)$.

(iii) In each operating period the company collects a constant amount of premium P_n=P.

(iv) If at the end of an operating period the company's capital exceeds Z the excess is paid out - as dividend or taxes.

(v) If at the end of an operating period the capital is negative, the company is ruined, and has to go out of business."

The proposed model is actually a random walk with an absorbing barrier for the accumulated Surplus (S) at S=0 and a reflecting barrier at S=Z (the predefined level of action).

(b) **Comments**: It is obvious that if we let Z (the reflecting barrier) to go up to infinity the model coincides with Lundberg's approach. The current model describes in a better way the reality because it assumes the existence of an upper limit for the accumulated surplus.

It actually proposes a certain type of control action on the surplus procedure. This action is not a smooth control as we only intervene in the system when S=Z.

Smoother control actions are proposed in the next papers (see Martin-Löf (1983)) operating in a more frequent way.

Finally we must stress that De Finetti solved his model only for a special case where P=1 (premium equals the unity) and the distribution F has the following form:

$$F'(x)=\begin{cases} p\ , & x=0 \\ q\ , & x=2 \end{cases}\ ,\ p+q=1 \quad \text{(where } F'(x) \text{ is the first derivative of } F(x)$$

The full solution has been provided by Borch (1967) who expanded the current model in the most general form (see next section).

3.3 Borch (1967), "The theory of risk"

(a) **Description:** An instructive journey from the roots of traditional risk theory up to the modern developments is presented as an introductory historical note in the paper. The author traces the footprints of the risk concept and premium determination procedures from the well-known "principle of equivalence" with the possible existence of a "safety loading" up to more elegant views of the utility theory, the economic theory of insurance and the most recent one, the control theory approach. It is suggested that the traditional view of adding a certain safety loading to the net premium and allowing completely free the reserve process of the company there after, creates an unrealistic model of a random walk with a positive drift going up to infinity. That's why, De Finneti's model is proposed again with an absorbing barrier at S=0 and a reflecting barrier at S=Z (S is the reserve or the accumulated surplus and Z a predefined level of action for the reserve). Finally the general solution of De Finetti's model is provided with some numerical illustrations.

(b) **Comments:** The current paper along with the previous one, provides the initial theoretical strike which "passes the ball" to the field of control theory approach. The modelling structure fits very well to a control problem. The existence of the reflecting barrier (S=Z) implies a certain control law (non-linear) which simulates the reality in a better way than traditional approaches. Apart the complex type of equations (Fredholm's type) and the respective complicated methods (iterated kernels, Dirac function, Neumann's expansion) used for the general solution of the model, the paper contains some very prudent observations either for the assumptions or the formulation of an insurance problem.

Firstly, indicates the most unrealistic assumptions which are being used in most insurance models i.e.

(i) The stationarity, which implies the nature of the company's business will never change.

(ii) The assumption that probability laws governing the process are completely known.

(iii) The implicit assumption that a decision once it has been made cannot be changed.

If we erase the above unrealistic assumptions then we shall be forced to consider a control approach to our model (a control model does not need these three assumptions).

Secondly, the paper eliminates the problem of an insurance company establishing

(i) An information system: a system for observing the stochastic process as it is being developed.

(ii) A decision function: a set of rules for translating the observations into action.

Especially, the last two rules will be widely used in the development of our applications.

3.4. Seal (1970) "Simulation of the ruin potential of nonlife insurance companies"

(a) **Description**: The paper suggests two types of modelling for a casualty insurance company calculating the probability or ruin using standard simulation techniques. The first model "consists of two independent and unchanging probability distributions. The first of these is the distribution of intervals between successive claims and the second is the distribution of individual claim amounts". Using queuing theory notation, the paper examines M/M/1, M/XI/1, XI/XI/1 situations where M stands for exponential and XI for Pareto. Finally it is observed that the longer tail distributions of interclaim periods and claims amounts respectively the bigger the probability or ruin.

The second model considers that the aggregate annual claim outgo of the company has a gamma distribution and different experience-rating strategies are examined calculating the probability of ruin. The results of the simulation establish a great surprise (paradox). Most of the simulated companies, although commenced with a fairly substantial reserve, during the first 40 time periods are ruined! The above, implies that standard rate-making strategies provide poor protection against the adverse chance fluctuations.

(b) **Comments**: The paper does not use control techniques but may contribute in two interesting directions (the second direction along with the two previous papers is the commencement point of the control approach to actuarial problems).

(i) The simulation of ruin potential of an insurance business especially for short term periods. Of course simulation techniques always provide a limited view but it is preferable than nothing. As we know ruin theory provides the formula which calculates the exact probability of ruin. But sometimes, it is very difficult or even impossible to obtain the respective numerical value.

The paper should be also considered as a starting point to a more general field of approximating the ruin functions.

(ii) The second direction which appears to be more interesting and stands as an extra incentive for this thesis is the result for the ruin potential of companies operating under "reasonable" experience rating procedures. We should pay more attention in the construction of these rate-making rules and be able to examine how much reasonable, effective or safe are, with respect to the ruin potential of a company. Actually the rules mentioned above should be considered as controllers and the respective premium formulae should form the basic equations of a dynamic system. Obviously the whole discussion above suggests the use of control theory which is the suitable tool for examining the rate-making rules.

3.5 Balzer & Benjamin (1980), "Dynamic response of insurance systems with delayed profit/loss sharing feedback to isolated unpredicted claims"

(a) **Description:** The paper considers a problem of group profit-sharing schemes providing a limited standard solution which is common to actuaries. It also uses that problem as an incentive in order to enter the field of control theory. A dynamic model is constructed, which simulates the actual problem with a delayed profit/loss sharing feedback mechanism. The basic equations are:

$$S_n = S_{n-1} + eP_n - C_n \tag{3.5.1}$$

$$P_n = \frac{1}{e}C_{n-f-1} - \varepsilon S_{n-f} \tag{3.5.2}$$

where S_n: accumulated surplus at time n,

P_n: premium received in (n-1,n)

C_n: claims incurred in (n-1,n)

e: expense factor

ε: feedback factor

The delay factor (f) is not fully examined. The authors consider the special cases for f=1,2 and f=5 in order to judge that increasing delays cause undesirable oscillations to the system.

The first equation describes the development of this accumulated surplus while the second operates as a decision function for premium determination using the latest known claim experience C_{n-f-1} and the latest known state of surplus S_{n-f} (f may be 0,1,2,3,...). The paper then concentrates to the dynamic behavior of the system after applying a disturbance input of an isolated unpredicted claim i.e. a spike input signal. It is proved that increasing delays cause undesirable oscillations which may ultimately lead to instability. Finally we must mention that, paper also outlines the modelling philosophy of a dynamic system and the perspectives of such a design for the insurance systems.

(b) **Comments:** An introductory paper for actuaries entering the field of control theory. Although the general concept of "control" have been included in previous research papers (e.g. Borch, (1967)) the specific approach of Balzer and Benjamin formulates the insurance problem under the typical notation, terminology and standard manipulation procedures of control theory. The reference to the modelling philosophy also confirms the statement above.

As regard the actual problem of group business, we may say that it is an ideal one in order to discuss the basic control concepts of input, output state and feedback mechanisms.

The authors also use control diagrams, a matter of great importance as the abstract insurance objects may find some kind of graphical representation.

The underlying philosophy of the system of equations (3.5.1) and (3.5.2) may comply with (while extending) the work and suggestions of De Finetti (1957) and Borch (1967). The control law is now smoother as we intervene into the system not only at the predefined level of surplus (Z) but at any time proportional to the volume of surplus.

The authors concetrate on the feedback concept and equation (3.5.1) which describes the state of the system. They discuss the stability of the model and the transient responses with respect to the feedback factor ε (identifying oscillations for large values but ultimate return to initial steady state).

The inconvenient point of the discussion is the use of a poor claim predictor. As we can see from equation (3.5.2) the experience of (n-f-1)th year is only used instead of an average of the last p-years. The last more sophisticated predictor may have reduce the oscillations of the feedback.

The model and research of this paper is being extended in the next chapter 4. The basic directions of extension are:

(1) the consideration of an investment element

(2) the introduction and investigation of the delay factor (f) as a free changing variable in the system.

(more details for the extension are provided in section (4.1)).

3.6 Balzer (1982), "Control of insurance systems with delayed profit/loss sharing feedback and persisting unpredicted claims"

(a) **Description:** At the beginning of the paper there is a short introduction by S. Benjamin who describes a "personal journey". He actually reports the practical questions or past research work (as the paper of Seal (1970)) which have directed him to a new conceptual universe using control theory in insurance problems. Then Balzer referring to the previous joint paper with Benjamin in 1980, attempts to formulate the same problem of profit/loss - sharing in a more sober way with respect to the control design and justification.

The system is disturbed by a persisting series of unpredicted claims (i.e. the input signal is a step function) and its behavior is examined with respect to transient responses and stability of the ultimate state.

The author goes "deeper" in control concepts describing the powerful tool of root locus method and its applications for the stability of the system or the optimal design of the parameters involved.

Using the final value theorem (in order to obtain the steady state value of the system) he finds that the system does not return back to the initial state. He then discusses different control actions (as proportional, derivative, integral or a combination of them) which can "correct" the behavior of the system and return it back to the initial state.

(b) **Comments:** The current paper has a direct relationship with the previous joint one (Balzer & Benjamin (1980)). It presents a clear and sufficient (or even say complete) view of the basic control concepts. It could be used as an autonomous reference for those actuaries who want to develop a more comprehensive understanding of the area of modelling and handling insurance problems via control techniques. As regards the results of investigation of the system we may point out the following.

(i) The important result is the determination of the optimal (with respect to the fastest response and return of the system back to initial state) value for the feedback factor which is equal to 31.25% (for a 20% cost factor). Of course, in a business context 31.25% appears as a very surprising proposal. A more "logical" and perhaps acceptable percentage (both from the insurer and the policyholder of the group profit-sharing scheme) should be a "round" one. Or in other words, as the certain percentage defines the way of sharing the profit, it appears fair to share 50%-50%.

(ii) Although integral action improves the dynamic behavior of the system, derivative action appears to degrade the dynamic response. The author provides an interesting general reasoning for this fact using the experience of driving a motor vehicle behind a large truck!.

Finally we should state again that no attention is paid to the claim predictor of the system and there is no discussion of the transient response characteristics of the system (i.e. to identify the basic concepts of section (2.18), delay time, peak time, rise time, maximum overshoot (see Section (2.18)).

3.7 Martin-Löf (1983), "Premium Control in an Insurance System an Approach using Linear Control Theory"

(a) **Description**: Using as a starting point the work of Borch (1967) the author concentrates on the premium control procedure of an insurance company. By a short review in the introductory section, he testifies the unrealistic view of the conventional risk theory but also indicates the weakness of De Finneti's model, in which the control action is restricted only at the reflecting barrier (S=Z). Consequently, he proposes a smoother control action which may stabilize the reserve in a more continuous way. Actually proposes a variable loading factor (θ_n) i.e.

$$\theta_n = f(S_n - Z) \tag{3.7.1}$$

where f is a function of the difference between the actual reserve at time t and Z the predefined level.

Following Borch's suggestions, he proposes a system of two equations in order to analyze the dynamic behavior of the company. The first equation describes the state of the reserve while the second determines the decisions about the premiums at the end (or beginning) of each time (operation) period. Upon the background above, the model is developed discussing the concepts of stability, ultimate state, and optimal selection of the parameters involved.

(b) **Comments:** In this paper, we can recognize the standard formulation of a typical dynamic system. The two basic equations which are similar to those of Balzer & Benjamin (1980) are:

$$S_n = RS_{n-1} + R^{\frac{1}{2}}(P_n - C_n) \tag{3.7.2}$$

$$P_n = \hat{\varepsilon}_1 \hat{C}_n - R^{\frac{1}{2}} \varepsilon S_{n-1} \tag{3.7.3}$$

where S_n: accumulated surplus at time n

P_n: premium received in (n-1,n)

C_n: claims paid in (n-1,n)

$\hat{C}_n$: claims predictor at time n

R: Interest factor

$\hat{\varepsilon}_1$: constant parameter (chosen arbitrarily)

ε: feedback factor

Removing the stationarity assumption, the author discuss a very sophisticated construction procedure for the claim predictor based upon linearity and random fluctuation which may actually simulate in a better way the real world. The discussion of the stability of the system with reference to the value of ε parameter is an easy task as the characteristic equation is a polynomial of 2^{nd} degree. Applying the theorems of section (2.15) the selection of parameter ε is restricted in order to obtain roots within the unit circle. The interesting (and rather expected) results is the discovery of a certain link between the feedback and oscillations of the system. (The same as in the previous section i.e. the stronger the feedback the bigger the oscillations.) Finally the proposed quantitative measure for the solidity of the steady state forms an optimal control strategy. The solidity requirements is given by the ratio

$$\text{Solidity Ratio} = \frac{\sqrt{\mathrm{var}(S)}}{E(S)}$$

which is restricted to a small value (e.g. 0.25) so that the probability of ruin is also restricted. This last rule returns back to the traditional approach of calculating the probability of ruin.

3.8 Benjamin (1984), "An Actuarial Layman looks at Control Theory"

(a) **Description:** The title of the paper explains in the best possible way the discussion and the effort made by the author to put actuarial problems into the control theoretical framework. An extensive analysis is provided in order to formulate a similar model as with the sections (3.5) and (3.6) paying much attention on the definitions of input, state and output variables. The model now is adjusted to a general insurance company. The basic concepts of traditional control theory appears to be adequate enough, for the simple model of single input-single output. The system is also subject to a ramp input signal as well as to a stationary random one. The author provides another application of control theory, in the area of pensions based on the aggregate funding method. A pension fund of a simple structure which uses the last funding method is considered as a control system. The investment return simulates the input while the fund level describes the state and the contribution rate represents the output of the system. The model is non-linear and provides an excellent opportunity to work with linearization methods while showing the potential application of control theory to the area of pension funds.

(b) **Comments:** A paper which should be placed in line with the papers in sections (3.5), (3.6). We must recognize in it some new efforts as regards the better and more elegant control modelling i.e. The author attempts to explain clearly why claims should be considered as input and surplus as the output variables while describing as state of the system each pair of input and respective output of the system.

The above definition of the "state" does not comply with the standard terminology and conceptual design of modern control theory. As we have seen in section (2.6) the "state" of the system in this paper is a vector of two consecutive values of the accumulated surplus while the premium is designed as output variable. The manipulation of the stationary random input signal is also a new kind of work which incorporates the stochastic element in the whole discussion of deterministic signals.

Finally, we must comment on the second non-linear model. The linearization method is not formal but results in the right equations (the formal framework is provided in section (2.12)). Although the designed model stays far away from the real structure of a pension fund it is a contribution to the effort of enhancing the possible applications of control theory into other actuarial problems.

3.9. Pesonen (1984), "Optimal Reinsurances"

(a) Description: The paper starts with the realization that any insurance company would prefer an optimal (in some sense) reinsurance arrangement. Under this fact, the author tries to answer the main question: "what kind of reinsurance treaties can be optimal?"

He reports the traditional treaties concluding that any optimal reinsurance arrangement "should be of function form i.e. the shares of the companies should be uniquely determined by the observed aggregate claims amount". Optimality is examined under the light of (concave) utility functions or using some other criteria assuming unknown utility functions (which normally is the case).

(b) Comments: A very mathematical paper which concentrates in the question of optimal reinsurance. Using quite complicated methodology and the respective theorems and proofs arrive at the optimal family of reinsurance treaties. The use of control theory is limited (and only) along the basic way of thinking and modelling structure.

3.10 Rantala (1986), "Experience Rating of ARIMA processes by the Kalman Filter"

(a) Description: The paper concentrates on the experience rating procedure of a claim process with a certain structure (ARIMA, a stochastic process defined by (3) three figures p,d,q where in the special case p=d=q=0 we obtain the white noise process). The basic target is the optimization of the sequence of proposed premiums restricting the variation of the accumulated profits. Optimization is defined in terms of smoothness of the premium flow into the system. Minimizing quadratic performance criteria and using the Kalman filter technique the author arrives at the optimal rules for premium determination.

(b) Comments: The author uses a very similar (with previous sections) equation to describe the development of solvency of the system i.e.

$$S_{n+1} = R_n S_n + P_n - C_n \tag{3.10.1}$$

(all the symbols have already been defined before). Then assuming a quadratic performance index which weights the variance of the solvency towards the d^{th} difference of premiums and the Kalman filter he determines the optimal premium rating strategy. The paper is a fine contribution to stochastic modern control theory as all the equations, variables and criteria are "transformed into the state-space form".

3.11 O' Brien (1987), "A two-parameters family of pension contribution functions and stochastic optimization"

(a) Description: The author starts from the fund ratio (fund value over the present value of the future benefits) as defined in Trowbridge (1963). He then models a pension fund with

(a) a controlled diffusion process and

(b) a quadratic performance index weighting the input variable (contribution rate) towards to the deviation of the fund from the suggested level of funding according to the fund ratio mentioned above. Finally he obtains the optimal control function for the contribution rate using Bellman's optimality principle. The author does not assume a target contribution rate as Vanderbroeck (1990) does in her approach.

(b) Comments: The paper uses advanced stochastic control theory (in continuous time) to approach the problem of pension funding. Setting a target fund ratio (in line with that defined by Trowbridge (1963)) the problem of funding may well be described as a typical control problem. The powerfulness of the approach relies on the advanced mathematics which can produce the analytical form of the optimal control function. From that form we can obtain deep insight into relations between the contribution rate and the known variables of benefits outgo (assumed linear), and the fund levels.

3.12 Taylor (1987), "Control of unfunded and partially funded systems of payments"

(a) **Description:** The paper considers unfunded systems (which maintain a fund approximately zero) and partially funded systems of payments (which maintain a non-zero fund at a level inadequate for full funding). Its target is the establishment of "a premium formula which is consistent with long term planning e.g. target rate of funding, limited variation in premiums from year to year etc.". Examples of such systems may be the Social Security Schemes of many countries (e.g. UK, USA, Canada, Greece).

The development of the model is based on the previous works of Balzer & Benjamin with the following modifications (generalizations).

(i) Introduction of an allowance for investment income as partial funded systems may well involve long term business.

(ii) A more generalized premium formula.

(iii) An explicit allowance for an initial value of the fund.

(iv) The rebate of premiums depends on the paid claims and not on the claims incurred.

As regards the second modification (i.e. the generalization of premium formula) the following general rule is used

$$P = P_b + P_c + P_p + P_s \tag{3.12.1}$$

where P_b represents a base (or target) premium which has been initially established while the other three premiums P_c,P_p,P_s represent adjustments made by the system evaluating the respective history of claims (C), past premium (P) and accumulated surplus (S).

Finally the results of the calculations indicate that instability increases in line with the delays of the system.

(b) **Comments:** This paper (considering the unfunded and partial funded systems) must be placed within the potential applications of the basic model developed by Balzer & Benzamin. The introduction of an investment component enhances our view and insight in the basic model, although there is not much weight on this item. The current model may well be combined with the works of Rantala (1988) and Loades (1992) (see sections (3.14) & (3.19) respectively) which refer to economic cycles and underwriting cycles. It will be an ideal vehicle in order to examine the behavior of the system under the simultaneous effect of investment and claim cycles.

The whole development of the paper is based on traditional control theory answering effectively the questions of the model since the basic structure of single input - single output is being used.

Finally we shall quote a remark of the author which appears in the ending part of the paper "Further Research".

"... The classical control theory used in this paper in common with that in the papers of Benjamin and Balzer is many years old. It is possible that the more powerful modern control theory, develop more recently could be used to advantage on insurance systems".

The last proposal of the author may be placed as the basic target of this thesis.

3.13 Rantala (1987), "On experience rating and optimal reinsurance"

(a) **Description:** This paper formulates a general reinsurance problem using a stochastic control theoretical model. Optimal control is obtained "by means of variances of such variables as underwriting result of the insurer, solvency margin or the insurer and reinsurer and the premium paid by policy holders".

The model is based in the following important assumptions as stated by the author:

(i) A reinsurance contract between two insurance companies (the cedant and reinsurer) has been made for a fairly long period and both parties will look for an arrangement which would be optimal (under some criterion) over a longer term (this assumption justifies among other things the use of asymptotic methods).

(ii) The reinsurer's annual share of the total claims amount is a function of present and past annual total claim amounts only (i.e. reinsurance does not depend on individual risk)

(iii) The reinsurer's share is a linear function.

Furthermore the basic dynamic system is described by the following equations:

$$S_{1,n} = R_1 \cdot S_{1,n-1} + P_{1,n} - C_{1,n} \tag{3.13.1}$$

$$S_{2,n} = R_2 \cdot S_{2,n-1} + P_{2,n} - C_{2,n} \tag{3.13.2}$$

$$P_{2,n} = P_n - P_{1,n} \tag{3.13.3}$$

$$C_{2,n} = C_n - C_{1,n} \tag{3.13.4}$$

where $S_{j,n}$ stands for the accumulated surplus at the end of year n for the j=1,2 company

$P_{j,n}$ stands for the premium received in year n for the j=1,2 company (P_n = total premium)

$C_{j,n}$ stands for the claims incurred in year n for the j=1,2 company (C_n: total claims)

R_j stands for the accumulation (interest) factor of the j=1,2 company

generally the subscript (1) represents the cedant while subscript (2) represents the reinsurer.

(b) **Comments:** The author attempts to provide a dynamic view of the general reinsurance problem improving the traditional static one. The most common approach is to split the total annual claims of a fixed year period (one year) into cedant's and reinsurer's part in an optimal way. This approach is extended to a longer time horizon in this paper.

The first version of the model considers "that $E(C_n)$ is known and both the total premium P_n and the reinsurer's premium $P_{2,n}$ are constants". Then the author searches for the optimal linear reinsurance policy i.e.

$$C_{1,n} = a_0 \cdot C_n + a_1 C_{n-1} + \dots \tag{3.13.5}$$

under the following restrictions.

(i) minimization of $SD(C_{1,n})$ when $SD(C_{2,n})$ is restricted to a given value (or vice versa)

(ii) minimization of $SD(\Delta C_{1,n})$ when $SD(\Delta C_{2,n})$ is restricted to a given value (or vice versa).

where SD(...) denotes Standard Deviation and

$$\Delta C_{1,n} = C_{1,n} - C_{1,n-1}$$

The second version of the model relaxes the assumption of constant total premium P_n and reinsurer' s premium $P_{2,n}$ and calculates the optimal control with respect to the available variables.

Finally as the author states the current paper "is more to show a feasible way to attack the problems of reinsurance than to give explicit results directly applicable in practice".

3.14 Rantala (1988), "Fluctuations in Insurance Business Results Some Control Theoretical Aspects"

(a) **Description:** The paper starts with a realization that, "The latest downward phases of underwriting cycles have been unusually deep and longstanding" and consequently proposes "proper analysis and modelling of fluctuation along the lines of:

(i) effective management

(ii) supervision and control of their solvency

(iii) fiscal purposes

(iv) reinsurance cover

(v) scientific premium rating".

Although there are conventional tools for actuaries to handle the problem above the author suggests control theory techniques as the new powerful and more efficient way of modelling.

The paper may be separated in the following four parts (with respect to the subjects discussed). In the first part, empirical data are presented and the respective conclusions are drawn such as:

(i) Loss ratios often exhibit cycles with a length of 4-8 year while amplitude of other characteristic may vary greatly from country to country or line to line.

(ii) Solvency ratios exhibit cycles with two or three-fold amplitudes because of accumulation of profit or losses, such cycles are crucial for financial strength.

In the second part the potential background for the cycle fluctuations is identified among the time delays, the forecasting techniques, faults in the tariff structure, fluctuation in investment return etc.

In the third part, the control theoretical tools are being developed using similar dynamic equations as with those of Balzer & Benjamin (1980).

Actually the first equation which describes the accumulation process is the following

$$S_n = R \cdot S_{n-1} + P_n - C_n \quad , \tag{3.14.1}$$

The differences (of the equation above) with equation (3.5.1) are the following:

(i) It includes the Accumulation factor R.

(ii) It does not include the expense factor e.

As regards the second equation which describes the premium decision formula is the following

$$P_n = P_{tg} + \hat{\varepsilon}_2 \cdot \left(S_0 - S_{n-2} \right) \tag{3.14.2}$$

where P_{tg} is a target premium while

S_0 is a target surplus level and

$\hat{\varepsilon}_2$ proportion similar to feedback factor (ε).

The equation above differs from equation (3.5.2) in two points:

(i) It does not use a claim predictor but a target premium P_{tg}.

(ii) It uses a target surplus level in order to decide the feedback action.

Finally in the fourth part the results of the formulation above are presented along with some more empirical data.

(b) **Comments:** The central and of course the very interesting concept of this paper is the existence of the underwriting cycle. The author suggests that this cycle has similarities with the business cycle although it is not properly defined and can not be easily identified from the official reports as the smoothing procedure hides the phenomenon (totally or in some extent).

The existence of the cycle is the best practical answer why control theory should be used in insurance problems. At this point we may also agree with the author's opinion that a claim process should be modelled as a combination of impulses, steps, ramps (trends) sinewaves (cycles) and other stochastic items.

An interesting view of the author is the construction of a general premium formula

$$P_n = P_{tg} + \sum_{j \geq 0} a_j \left[S_0 - S_{n-j}\right] + \sum_{j \geq 0} b_j C_{n-j} + \tau_n$$

where a_j and b_j $j \geq 0$, are weighting factors while τ_n is a correction factor compensating the stochastic components of the claim process. The first summation may remind us the integral action while the second summation represents a sophisticated claim predictor. The summation in the equation above assumes the use of the maximum available information. By general reasoning we could say that the more information will create less "problems" in the process (i.e. less fluctuating oscillation). The introduction of τ_n produces even better results canceling the undesirable fluctuations of random events.

3.15 Benjamin (1989), "Driving the Pension Fund"

(a) **Description:** The paper aims to present a short introductory note for the use of control theory in pension funding methods. Actually, it considers two of them: the "aggregate" method and the "projected unit" method. The input signal is the earned real rate of interest each year while the output signal is the recommended contribution rate. In order to obtain a direct comparison of the two methods, the same model fund (as regards the demogratic assumptions and the benefit structure) is used. Both of the two methods produce non-linear equations consequently the linearization method is used again in order to obtain the linear standard format of a dynamic system.

The input signal (real rate of return) is modelled using two deterministic functions (spike, step) and one stationary random one. The simulation results and diagrams are obtained while the author focuses on the ratio of standard deviations of input and output signals i.e.

$$\text{Ratio} = \frac{\text{SD}(c_n)}{\text{SD}(i_n)}$$

where $\text{SD}(c_n)$ standard deviation of contribution rate and

$\text{SD}(i_n)$ standard deviation rate of return.

The final part of the paper deals with an alternative proposal of controlling a pension fund. Actually a fund is driven from some initial conditions of fund level and contribution rate (F_0, c_0) to the ultimate (target) ones.

(b) **Comments:** The author manages to formulate another funding method (projected unit) under the conceptual framework of control theory. Under the achievement considered above, the formulation of the remaining funding methods (actually the exploration of the control characteristics of each one) may be just a matter of time and careful reference to the existing analysis.

The most interesting thing of the paper is the final part which actually appears in the title i.e. "Driving the Pension Fund". The author assumes the very practical problem of drawing a smooth path for contributions in order to achieve a target state (for contribution and fund level). The requirement of the derivation of "the smoothest path" is translated to an optimal control problem known as "minimum energy". The restriction of the time-horizon (4 years) of the problem enables the detection of a solution using Lagrange multipliers and some "hand writing" work. The possible extension of the time horizon for several years (e.g. 10 or 20) or the introduction of other additional performance criteria to the problem will lead to more complicated systems of equations which may demand heavy computer applications or approximation procedures. Such an extension will be considered later in section (6.15) where similar development will be used for the PAYG funding method.

In chapter 6 of our thesis, we shall use the modelling philosophy of Benjamin to formulate the PAYG (Pay-As-You-Go) funding method under the framework of control theory. We have chosen that method (for some reasons explained in chapter 6 and) in order to enhance the field of applications not only in the fully funded systems but also in the unfunded (or partially funded) ones.

3.16 Vanderbroeck (1990), "Pension funding and Optimal Control"

(a) **Description:** The paper considers a deterministic model for a defined benefit pension plan and by the means of optimal control theory defines the solution for the optimal contribution rate c(t) and the target contribution rate c^*.

The basic equation which describes the fund level (or the state of the system is the following

$$F'(t)=\delta\cdot F(t)+c(t)\cdot W(t)-B(t) \tag{3.16.1}$$

where each variable depends on time t. F(t) stands for fund level, c(t) for contribution rate and B(t) for benefits at time t while δ is the force of interest. The other important expression which should be minimized is the following:

$$\min_{\rho,\pi} \int_0^T e^{-\varphi t}\left\{[c(t)-c*W(t)]^2+\beta[\pi(BN)(t)-F(t)]^2\right\}dt \tag{3.16.2}$$

where	W(t)	total salaries at time t
	(BN)(t)	present value at time t of the future benefit for active and retired members
	φ	discount rate
	π	the required funding level as proposed by Trowbridge (1963)

The paper uses advanced optimization methods from functional analysis considering the Hamiltonian of the system and obtaining the optimal solution for c after determining a constant level for parameter c^*.

Finally the theoretical model is applied to Belgian Social Security System using linear functions for the future projection of the Benefits and the salaries up to the year 2050. Consequently, numerical results are obtained which from the sensitivity analysis seem to be reliable for the 30 years.

(b) **Comments:** The paper is a fine contribution to the potential applicability of optimal control into pension funding (in line with that of O' Brien (1987), see section (3.11)). It uses the most advanced tools of the functional optimization theory in order to obtain the solution in the analytical form while leaving the model in the most general continuous format. It will be interesting to compare and investigate a possible relationship and similarities of the current paper with the work of Benjamin (1989).

Actually the two problems are very similar in the sense that both of the two models target the optimal path of contribution rates. Benjamin (1989) reduces the complexity of the problem by considering it in a discrete time context, with a limited time horizon (4 years) so the optimization procedure is restricted to the use of Lagrange multipliers. The continuous format of the current paper forces the analysis to advanced functional minima methods. As regards the theoretical result we must stress again the complexity but also the elegant format which ensures insight into the problem.

Finally, we should stress the link of this paper with our work in chapter 6.

3.17 Brown & Grenfell (1992), "Crediting Rate Management for Capital Guaranteed Insurance"

(a) **Description:** As is stated in the introduction crediting rate management attempts to satisfy multiple objectives such as:

(i) Solvency

(ii) Investment Return

(iii) Marketing

(iv) Financing Reserves for normal and abnormal growth

(v) Objective crediting rate formula

(vi) Equity

The complicated problem above is tackled in this paper using historical data and simulation studies supported by models of control theory. The basic equation of the model is similar with that of Rantala (1988) (see equation (3.14.2)) for premium determination. Now instead of charged premium P the target Premium P_{tg}

and target surplus level we have Adjusted Earnings Rate (AER), Clients Expectation Indicator Rate (CEIR) and Target Interest Equalization Reserve (TIER).

$$(\mathrm{AER})_n = (\mathrm{CEIR})_n(1-t') + \hat{\varepsilon}_3\left((\mathrm{AIER})_n - (\mathrm{TIER})_n\right) \tag{3.17.1}$$

where t': allowance for tax

$\hat{\varepsilon}_3$: feedback factor

AIER: Actual Interest Equalization Reserve.

The feedback factor is designed in a special way

$$\hat{\varepsilon}_3 = \begin{cases} \hat{\varepsilon}_{3,1} & \text{if} \quad (\mathrm{AIER}) < (\mathrm{TIER}) \\ \hat{\varepsilon}_{3,2} & \text{if} \quad (\mathrm{AIER}) > (\mathrm{TIER}) \end{cases} \tag{3.17.2}$$

with $\hat{\varepsilon}_{3,1} > \hat{\varepsilon}_{3,2}$ (3.17.3)

(i.e. the feedback action is reduced when the (AIER) is greater than (TIER). "The developed simulation model was used to test the effect of changes in the feedback and (TIER) parameter on various quantities over a simulated period of ten years".

(b) **Comments:** A very interesting paper as it transfers the basic model of Rantala (1988) into the area of "Rate management". The author relies heavily on simulation results constructing the respective tables and identifying the optimal pair of values (feedback factor and TIER) according to some given criteria.

Generally speaking the feedback factor operates parallel with (TIER) i.e. For a certain quantity the increase of feedback factor will produce analogous results with the increase of (TIER).

The idea of an asymmetric feedback factor is very impressive. Up to now all the papers (we have examined) assume a constant feedback factor. The new approach adds an interesting flexibility in the model design. Further improvements may be achieved by considering a more complicated function for the feedback factor.

Another interesting direction of research is the proposal of a feedback form as

$$\hat{\varepsilon}_3(\log(1+(\mathrm{AIER}))-\log(1+(\mathrm{TIER})))$$

Because as the author states:

"This form gives greater proportional correction when the (AIER) is below target for a fixed value of the factor $\hat{\varepsilon}_3$".

3.18 Kamano & Nara (1992), "Long term management of Reserve & Dividend"

(a) **Description:** The current paper is involved with a very similar subject as with the previous one of Brown & Grenfell (1992) i.e. the management of the reserve and dividend in a mutual life insurance company. The problem is converted to the equivalent one i.e.: Describe the investment behavior of the company and determine the procedure of splitting the rate of return in two parts, a steady one which should be distributed to policy holders and a fluctuating one which should be retained as a reserve. For this reason, the following equation is proposed

$$j(t)=j_1(t)+j_2(t) \tag{3.18.1}$$

where $j(t)$ = rate of return of a stock during year t

$j_1(t)$ = rate of (steady part) return

$j_2(t)$ = rate of (fluctuating part) return

$j_2(t)$ can be regarded as white noise, while $j_1(t)$ may be assumed to perform a random walk. The equations above along with the following one

$$F(t)=(F(t-1)+(CF)(t))\,(1+j_1(t)+k(t)\cdot j_2(t)) \tag{3.18.2}$$

which determines the consecutive asset values F(t) formulates the dynamic equations of the model (CF)(t) is the cash flow in year t and k(t) is the proportion of fluctuating asset in year t).

The solution of the problem is obtained using the powerful tool of "Kalman Filter". (see Rantala (1986)). The disussion is completed with the establishment of the control parameters.

(b) **Comments:** The modelling procedure of the paper is interesting as it proposes the use of the "Kalman Filter" for the solution of the problem which appears to be a promising tool for potential application in insurance problems.

The unique characteristic of this paper is the absence of a feedback factor. Of course, the feedback mechanism may be considered as a hidden element of the filtering procedure.

The formulation of the investment behavior as a random walk ensures the suitability into the filtering algorithm but restricts the potential use of other different patterns as an input signal in the model.

3.19 Loades (1992) "Instability in Pension Funding"

(a) **Description:** The paper presents the most general type of actuarial model which may be used for the controlling of a pension fund. The basic parts of this model as described in the paper are:

(i) Benefit structure.

(ii) The valuation method which determines how the actuarial liabilities and standard contribution rates are to be calculated.

(iii) Demographic assumptions.

(iv) A model of market rates of return.

(v) A decision rule to produce valuation rates of interest from the investment model.

(vi) A method of placing an actuarial value to the fund.

(vii) A control mechanism for amortising surpluses / deficiencies.

The paper considers the entry age funding method and investigates (using a numerical approach) different control strategies (regarding some of the controllable parts mentioned above) with respect to the mean and variabilities of the actual recommended contribution rates (and the resulting funding levels and surpluses). The whole model is placed (and studied) a 15-year economic cycle which continuously repeats itself and completely determines the rates of investment return. For the reason above, exponential smoothing is used for setting the valuation rates of interest (and also because the last method seems to simulate better the observed behavior of the actuaries compared to the arithmetical smoothing method). Finally, numerical results are obtained and discussed with respect to each parameter involved.

(b) **Comments:** The central concept in this paper, which eventually has influenced the analysis and the respective results is the principle of a 15-year economic cycle. This idea may well correspond with the concept of the underwriting cycle (with a length of 4-8 years) developed by Rantala (1988). It will be very interesting to explore the possible relationships between them, finding intersections, parallel movements etc.

The results of these two papers indicate at least similarities in the generation of cyclical oscillations in the output variables of the two models. By general reasoning we may argue that the coexistence of the two cycles will produce multiple effects (if parallel) or will cancel any oscillation (if opposite). As regards the other results of the paper we should emphasize that ultimate position is never reached due to the heavy influence of the transient effects which may be attributable to the exponential smoothing procedure (which actually operates as a claim predictor in the model).

All the above observations may drive us to the key conclusion of the author that "If investment returns are driven by a cyclical economy, static valuation bases reduce fluctuations in both contributions and surpluses".

Finally, as regards the use of control theory in the development of the paper we must say that it is restricted only in the general philosophy of "control" and the basic concepts of the theory.

3.20 Martin-Löf (1994) "Lectures on the use of Control Theory in Insurance"

(a) **Description:** The paper discusses the theoretical background of the optimal control techniques in Markov chains and some practical applications to non-life insurance business. The definitions of the minimal cost and the respective equation which holds between the states i.e.

$$V(i)=\min_{u}\left(c(i,u)+\upsilon\sum_{j=1}^{N}P(i,j,u)\cdot V(j)\right) \tag{3.20.1}$$

where V(i): minimal cost for the i state

c(i,u): operational cost at state i and applying the u control

P(i,j,u): Probabilities, $P(i,j,u)\geq 0$ and $\sum_{i=1}^{N}P(i,j,u)=1$

υ: discount factor

N: total number of possible states of the system

play the most important role in the development of the models.

The basic equation (3.20.1) is used in order to handle De Finetti's model minimizing the present value of all future premiums. The solution is provided for the special case of a random walk with absorption and reflection barriers. The model of Martin Lőf (1983) is also discussed applying two different input signals: a stationary stochastic variable and a deterministic sequence of a "step" signal.

(b) **Comments:** The author uses a simple and self-contained terminology and reasoning in order to present an introductory survey on optimal control of Markov chains for Actuaries with no previous training in that area.

In the solution of De Finetti' s model, he calculates the optimal control, finding the optimal reserve level for the reflecting barrier ξ^*. Then calculates the probability of ruin and consequently check whether it stands below an acceptable level. A reverse approach might also be interesting i.e. firstly calculate the optimum probability of ruin determining the optimal ξ^* and then check the present value of all future premiums. Of course a weighed criterion (a compromise) for the optimal value of ξ^* may be constructed upon the two concepts described above.

The use of the Markovian control

$$P_n = P_n(S_{n-1}) \tag{3.20.2}$$

may not always applicable because the insurance business can exhibit longer delays. (It is common to have available values for surplus 2,3 or... years before i.e. S_{n-2}, S_{n-3},...). So it would be more appropriate to consider control actions based on S_{n-f} where f=2,3,... (this has been done by Balzer & Benjamin).

Another interesting feature of the paper is the linearization procedure around the equilibrium point for the general model of De Finetti. Using that method, we might calculate $E\{S_n\}$, $Var\{S_n\}$ and observe, the results a growing $E\{S_n\}$ which is not desirable. Hence, a compromise should be obtained by keeping $\{S_n, n = 1,2,...\}$ reasonably small while averaging $\{P_n, n = 1,2,...\}$.

3.21 Fujiki (1994), "Pension Fund Valuation"

(a) **Description**: The author presents a general overview of the "valuation" process (along with other aspects of the required management procedure) of a pension fund. After providing the necessary theoretical background for pension funds he focuses on the potential applications of control theory into this area. He investigates the control of the valuation basis through a changing environment with respect to

- movements in the real investment returns
- movements in the dividend yield and
- movements in the rate of withdrawals.

He proposes certain patterns for the movements of the parameters above and carries out projection for several set of numerical values.

(b) **Comments**: The author uses the basic philosophy and elementary concepts of control modelling. He provides some theoretical background for open-loop and closed-loop systems (with special reference to feedback action) and also some more details about the concepts of input state and output variables of a dynamic system. He then investigates the "pension fund valuation" process under several different scenarios with respect to the input signals (actual experience follows the patterns of a spike a step or a sine wave) and the possible control strategies for the selection of the valuation basis (constant or changing patterns). Finally he proposes an optimal selection procedure using averaging or past experience and/or small delays in the decision formula for the valuation basis.

3.22 Haberman & Sung (1994), "Dynamic approaches to pension funding"

(a) Description: In the introductory section of the paper the authors describe the basic philosophy, structure and mechanisms of an occupational pension scheme while pointing out that the management of a pension fund is an excellent example of a control problem. Having established the control characteristics of the problem they discuss the conflict of interest between the employer and the trustees / employees. The former aims to stability while the last aims to security. This conflict of interest is converted into a functional which weights the "contribution rate risk" (stability) towards the "solvency risk" (security). The functional is minimized under different scenarios obtaining optimal control for the premiums.

(b) Comments: The paper is a fine contribution to the modelling of a pension fund using optimal control techniques. It discusses both the deterministic and the stochastic versions of the problem. The form of the functional used is quite general and actually extends the different formats proposed by Benjamin (1984), O' Brien (1987) and Vanderbroeck (1990). The functional has the form (the discrete version)

$$J_T = E\left\{ \sum_{t=s}^{T-1} \upsilon^t \left[(c_t - \tau c_t)^2 + \upsilon \cdot \beta \cdot (F_{t+1} - \tau F_{t+1})^2 \right] \right\}$$

F_t : fund level at time t

τF_t : target fund level at time t

c_t : contribution rate in year (t+1)

τc_t : target contribution rate in year (t+1)

υ : $(1+i)^{-1}$, i valuation real rate of return

β : a weighting factor to reflect the relative importance of the solvency risk against the contribution rate risk

(the continuous version requires integral instead of summation).

Having established the objective function and the control horizon the paper obtains the optimal control $c_s, c_{s+1}, \ldots, c_{T-1}$.

3.23 Sung (1997), "Dynamic Programming Approaches to Pension Funding"

(a) **Description**: As the title of the thesis indicates, the author uses a dynamic programming approach to pension funding. Setting up a dynamic model he uses optimal control techniques to balance the conflicting interests of the employer and the trustees or employees for stability and security respectively. He uses two distinct linear models based on different forms of the solvency equation.

[(a) the "modified" solvency-level growth equation and
(b) the "zero input, 100% - target solvency-level growth equation".]

Finally he derives "optimal funding control procedures for the contribution rate by solving" seven special versions of the control problem.

(b) **Comments**: It is a contribution to the application of control theory (especially the dynamic programming) into the actuarial problems (especially for pension funds). The formulation of the problem follows exactly the rules and terminology of control theory with emphasis on dynamic programming methods.

One of the basic concerns is the optimization of the "spreading parameter" which becomes extremely difficult when there is a large number of boundary constraits and long control horizon.

Another issue of the research work is the comparison between the "dynamic pension funding" and the "spread funding" plans. The author proposes the following equation

$$DC_t = NC_t + pc_t UV_t + ac_t$$

as a more suitable formula in order to balance the conflicting interest of the employer and trustees or employees, where

DC_t: Dynamic Contribution rate applying between time t akd t+1

NC_t: Normal Cost applying between time t and t+1

UV_t: Undesirable Valuation outcome at time t (e.g. AL_t-F_t for a classical actuarial valuation)

pc_t: proportional controlling parameter applying between t and t+1

ac_t: additive controlling parameter applying between t and t+1

3.24 Loades (1998), "Elementary Engineering Control and Pension Funding"

(a) Description: This paper, in line with previous works Benjamin (1984), (1989) and Loades (1992), explores the relationship of the control theory with the area of pension funds. The author uses a "continuous type" model to describe the development of the fund level (F_t) assuming also a "proportional controller" to feed back (and control) the unfunded liability into the system. In the actuarial literature the "proportional controller" is known as "spread forward method". Additionally, equations are modified to reflect the development of the solvency level of the scheme. Finally, derivative and integral control actions are introduced in order to eliminate the stable error and upgrade the efficiency of the whole system.

(b) Comments: The paper contributes in two directions. Firstly, it shows how integral action can improve the efficiency of a pension scheme under a control framework. Secondly, it investigates the behaviour of the fund (and solvency level) under an environment of cyclical rates of return. The last investigation shows that long cycles of 15 years may produce dramatic oscillations compared with rapid cycles of 5 years which may not affect the gradual development of the fund (and the solvency level).

3.25 Conclusions - Cross References to the papers

In this chapter we have collected and presented some of the most important papers which may well describe the points of intersection between the control theory and actuarial science.

ACTUAL PROBLEMS

As we have seen, applications cover all the area of actuarial work from life assurance and especially group business (section (3.5)), general insurance (section (3.20)), pension funding methods (section (3.15)), investment (section (3.17)), reinsurance (section (3.13)) and generally the concept of premium control and solvency requirement of an insurance company (section (3.7)) as well as probabilities of ruin (section (3.4)).

MODELLING PROCEDURE

As regards the modelling procedure, they all try to formulate the problems using pure control techniques providing also some limited solutions which may be more familiar to actuaries (section (3.5)). Most of the models use traditional or classical control theory as all of them are designed as

Single Input - Single Output

(although some of them are described with more than one input or output variables).

INPUT SIGNALS

In line with the standard control techniques most of the models are analyzed using some deterministic signals described with special names as: Spike, Step, ramp, sine wave.

Apart from the above ones, stochastic signals are also used measuring the variance of the output response (compared with the input's variance).

FEEDBACK ACTION - STABILITY

Feedback action is widely used in order to control the transient response or ultimate states of the system. The analysis of this action across all the papers show that feedback in conjunction with the inertia delay of the systems produces oscillations.

Special design should be considered for the choice of the parameters involved in order to avoid large undesirable oscillations and produce stable systems.

OPTIMAL CONTROL DESIGN

Many of the papers are involved with some kind of optimization procedure in the model. The first category discusses the optimal choice of the parameters which governs the stability of the system. Consequently, they search for the values which quickly return back the system to initial state. (The parameter which takes the major concern is the feedback factor).

The second category discuss the optimal choice of the parameters which minimize the variance of the output variable.

The third category is involved with the more advanced optimal control techniques trying to obtain the optimal function control of the system. Such examples are:

(i) In section (3.15) the determination of the optimal "smoothest" path for contributions and fund levels of a pension fund.

(ii) In section (3.16) the detemination of the optimal (which minimizes a certain functional) function as a choice for the development of the contribution rates in a pension plan operating under the "pay as you go" system.

STOCHASTIC / DETERMINISTIC APPROACH

Almost all the authors refer both to the deterministic and the stochastic approach of each problem, with more emphasis on the former one. In the stochastic version of a model the variables are substituted by their means while for the optimizaton procedure the respective variances are considered. The Kalman Filter is proved to be an efficient tool to handle the stochastic component of the investment procedure presented in Kamano & Nara (1992). Another contribution is that of Benjamin's (1984) where he tackles a random imput signal by considering the minimization of the ratio between (a) variance of the output response over (b) the variance of the input signal.

NON-LINEARITY

As the non-linear problems exhibit high degree of complexity there is a certain preference from the authors to use linear models. Of course, there are some of them (e.g. Benjamin (1984), Martin-Löf (1983), Vandebroek (1990)) who model the problem in the most general form (using non-linear equations) and then use the linearization procedure at the equilibrium point in order to obtain a solveable linear dynamic system.

MODERN CONTROL THEORY

Finally we must stress that all the papers identify the great potential of the use of modern control theory. Of course it is unavoidable to use it, if the model designed is of the form

multiple input - multiple output

or refers to a time-varying problem. The last area of research will be covered by the current thesis.

Chapter 4

Application to Insurance Pricing

4.1 Introduction

In this chapter, we shall design and develop a model of "Insurance Pricing" considering a very common problem which quite often causes "headaches" to insurance managers and underwriters, especially when it is placed in a wider context taking into account market competition. It will be formulated and solved using control theory techniques while providing standard (up to now) view of the actuarial approach.

The model is very similar to the one developed by Balzer and Benjamin (see sections (3.5), (3.6) & (3.8)). Actually we may consider it as an extension to that previous research work. This extension has been achieved by two ways:

(i) Incorporating additional modelling concepts and background from the similar papers of chapter 3 i.e. Firstly, we have introduced an investment element in the accumulation procedure (of the surplus process) as Martin-Löf (1983) and Taylor (1987).

Secondly we have improved the claim predictor considering the average of the last years using weights and inflation factors similar with the proposal of Martin-Löf (1983) and Rantala (1988).

(ii) Introducing entirely new concepts in the modelling design, trying to extent the model into the most possible directions e.g. Consider a time-varying interest factor, or leave the delay factor as a control parameter into the system and examine its affect in the solution.

Generally speaking, the target of this chapter is to provide a complete (as possible) reference to the "Insurance Pricing" techniques of the short term insurance policies starting from the roots of classical risk theory and "the principle of equivalence" up to the most modern view initially pointed by Borch (1967) and subsequently followed by many others as we have seen in chapter 3.

4.2 Description of the problem

We consider a non-life portfolio with available data from the past years. Let $C_1, C_2, \ldots, C_n$ denote the total amounts of incurred annual claims over the last n years

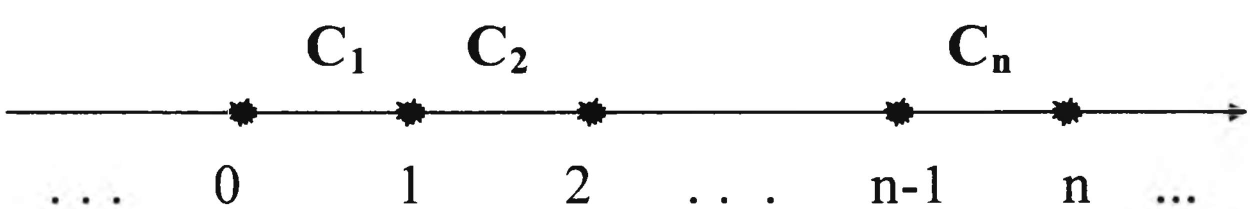


Diagram (4.2.1)

We are at the end of year n and we have to decide the premium for the next year. In the real world we may have to solve the above problem without knowing C_n, (the claim experience of the last year) or even C_{n-1} or C_{n-2}. And sometimes we may have no available data at all or any other similar experience. Of course the question is still there... "the next year's premium?".

But if it was only that question the problem could be easy. Another identical question arises exactly one year later!... at the end of year (n+1) which is the new premium for the next year (?), assuming all the recent experience including C_{n+1}. And after one year again, the same question and so on! Considering the situation described before we may quote Brian Hey's opinion from Balzer (1982): "It is all very well, trying to forecast the next few years but we don't even know what has happened in the last few years, we're still having to forecast that".

Finally, after all we realize there is a need to establish a formal decision procedure to determine the premiums for successive years. We have to build a control strategy to guide the whole system through a safe path.

It seems like an electronic game with a car. The player is a driver who tries to keep the car on the road using a wheel (turning left or right) while only a small part of the road is revealed in front of him on the screen.

4.3 Standard formulation and the respective notation

Let us try to give some quick answers to our "question" using the standard statistical - actuarial approach. Firstly we should standardize a formal notation for the purposes of our calculations.

- C_n, n=1,2,... total amount of (incurred) annual claims for the year 1,2,...(paid at the end of each year).
- P_n, n=1,2,... total gross (including expenses) premium for the year 1,2,... (paid at the beginning of each year).
- $1-e_n$, n=1,2,... total proportional expenses (including the desired profit margin) for the year 1,2,... expressed as a percentage of total gross annual premium. We also assume that expenses are paid at the beginning of each year and are constant, equal to 1-e without any fluctuation.

- L_n, n=1,2,... annual surplus at the end of the year n (i.e. at time n).
- S_n, n=1,2,... accumulated surplus at the end of the year n (i.e. at time n).
- R_n, n=1,2,... interest rate factor for the year n ($R_n=1+j_n$ where j_n is the real rate of return in year n).

4.4 First actuarial approach based on experience rating methods and/or the given (priori) distribution of claims

In the last section, we have established a lot of parameters to describe our problem. Now in order to get a first quick answer we shall remove some of them (equating with zero or other specific value) to handle better our calculations.

(i) $R_n=1$, n=1,2,... no interest rate (i.e. $j_n=0$ for every n).

(ii) Ignore annual and accumulated surplus variables.

Answer (a): Simple Average

Although we have found that the use of average function and generally the standard experience rating methods (Seal (1970), section (3.4)) may guide the insurance company into catastrophe, we shall provide it as the first obvious solution. An insurance manager with no other kind of information (except $C_1,C_2,...,C_n$) should go on and decide the next year's premium P_{n+1} by using the formulae below: (The symbol ^ above the claim parameter indicates an estimation for claims).

$$\hat{C}_{n+1} = \frac{1}{n}(C_1 + C_2 + ... + C_n) \tag{4.4.1}$$

$$P_{n+1} = \frac{1}{e}\hat{C}_{n+1} \tag{4.4.2}$$

Answer (b): Weighted Average

This answer requires more pieces of information which after a special interpretation may be reflected in the mathematical formulae of premium estimation. Here, we shall consider two specific examples.

1st Example: Assuming that C_n represents the annual claims for a motor insurer and government has imposed stricter traffic rules in the last years resulting a decline of claims, then we should weight more the recent experience.

Using exponential smoothing with higher weights for more recent claims i.e. $w_k=\theta^k$, k=1,2,...,n and $\theta>1$ we obtain, (where w_k, k=1,2,...,n are the weights for each observation C_n).

$$\hat{C}_{n+1} = \frac{\theta C_1 + \theta C_2 + \ldots + \theta^n C_n}{\theta + \theta^2 + \ldots + \theta^n} \qquad (4.4.3)$$

and again use equation (4.4.2) for P_{n+1}.

2nd Example: Assuming the situation of the 1st example plus the existence of an annual inflation factor, say $F>1$, which affects the development of the claim cost each subsequent year (i.e. if a certain car damage costs one unit of money for the current year, then the same one costs F the next year) we may use again the exponential smoothing technique. The weights now, reflect two corrections due to inflation, and the recent experience

i.e. $\qquad w_k = F^{n-k+1}\theta^k \qquad$ k=1,2,...,n

(4.4.4)

Hence $$\hat{C}_{n+1} = \frac{F^n\theta C_1 + F^{n-1}\theta^2 C_2 + \ldots + F\theta^n C_n}{F^n\theta + F^{n-1}\theta^2 + \ldots + F\theta^n} \qquad (4.4.5)$$

Again use equation (4.4.2) for P_{n+1}.

Answer (c): "The net premium principle (principle of equivalence)".

A standard method for premium calculation, which assigns the mean to each distribution of claims (see Gerber (1979)). Of course if we know the distribution of C_n's (e.g. $N(\mu,\sigma^2)$) and we absolutely believe it, there is no need to use the approach in answer (b) (i.e. evaluate the past experience). We may simply use the "principle of equivalence" and predict the claims using the mean μ i.e.

$$\hat{C}_{n+1} = E(\text{Claims}) = \mu \tag{4.4.6}$$

Again using (4.4.2) we derive P_{n+1}.

4.5 Second actuarial approach based on ruin theory (or probability of ruin)

The analysis of the last section is too simple. We have ignored all the variations which may occur and the surplus or deficit which arises due to trends, variations, fluctuations or bad premium estimations. In this section we shall focus on the surplus procedure and try to protect the company against the white noise (randomness of the events).

Answer (a): Traditional Ruin theory

The traditional ruin theory proposes a simple rule for premium determination. Claims (C_n, i=1,2,...) are supposed to be independent identically distributed random variables with a given distribution ($C_n \sim H(c)$, n=1,2,...). We find the mean μ of the distribution by two ways,

(i) directly, if we know the exact form of the distribution

(ii) indirectly, using conditional probabilities

$$\mu = E(C_{n+1} | C_1, C_2, \ldots, C_n) \tag{4.5.1}$$

Then, the risk premium is calculated adding a suitable safety loading θ_0. i.e.

$$\hat{C}_{n+1} = (1+\theta_0)\mu, \quad \forall n = 1,2,..., \tag{4.5.2}$$

and using (4.4.2) we obtain the premium P_{n+1}. Parameter θ_0 is chosen such that the risk of "ruin" i.e. $Pr(S_n<0)$ is sufficiently small (e.g. 10^{-2} or 10^{-3}). Using risk theory we can calculate the probability of ruin

$$Pr(S_n<0) \cong \exp(-R^* \cdot S_0) \tag{4.5.3}$$

at infinite time where R^* is the coefficient of ruin depending on the distribution of claims C_n, n=1,2,... and S_0 is the initial capital (for more analysis see Bowers et al (1980)).

The choice of the parameter θ_0 determines the name of the principle for premium estimation (see Gerber (1979)) i.e.

i) Expected value principle: θ_0 may take any value

ii) Variance principle: θ_0 is proportional to the ratio $\dfrac{Var(C)}{E(C)}$

iii) Standard deviation principle: θ_0 is proportional to the ratio $\dfrac{\sqrt{var(C)}}{E(C)}$

Answer (b): A slight modification to ruin theory

If we observe the accumulated surplus variable S_n, we should notice that it grows infinitely, since the risk premium has the positive safety loading ($\theta_0>0$). So a first obvious modification puts an upper bound S_ω to variable S_n. That means when S_n becomes greater than S_ω then the difference S_n-S_ω is distributed as a refund to policy holders. So the time diagram of surplus variable is bounded into one semi-plane as appears in the following diagram (4.5.1).

Of course, if the surplus S_n becomes less than zero then the company is ruined and the whole process stops. This model is due to De Finetti (1957) but also discussed by Borch (1967) who gave the general solution.

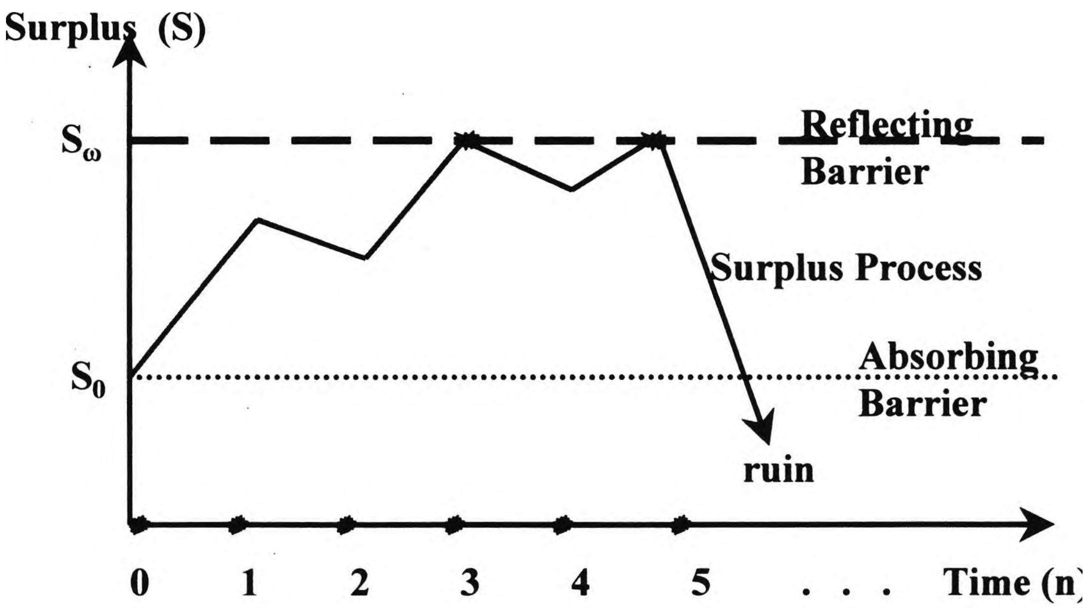


Diagram (4.5.1)

After this modification it turns out that the probability of ruin equals to one $(\Pr(S_n < 0) \cong 1)$ but when S_ω is a very large number the probability of ruin within a reasonable time period is given by the same formula as expression (4.5.3).

The results above are quoted in Borch (1967) and Martin-Löf (1983) (sections (3.3) & (3.7) respectively).

Answer (c): A more sophisticated version of ruin theory

There are many other versions of answer (a) and (b) which improve the behavior of the system. All the methods target to smooth premiums while keeping small the probability or ruin. The concept of "control strategy" is found in those versions of premium calculations.

For example a method which stabilizes S_n in a more continuous (uniform) way is given by the following formula

$$\hat{C}_{n+1} = (1 + \theta_0(S_n))\mu \tag{4.5.4}$$

The safety loading $\theta_0(S_n)$ depends upon the accumulated surplus and obviously decreases as S_n increases or vice versa $\theta_0(S_n)$ increases as S_n decreases.

The above procedure described by the formula (4.5.4) is due to Martin-Löf (1983).

4.6 Third actuarial approach based on "manual" control techniques

From the first and second actuarial approach, it is obvious that it is necessary to have some kind of control procedure in order to guide the whole system through a "safe path" as regards the smoothness of premiums, equity and solvency requirements.

The diagnosis of the necessity above, has been identified very early, almost from the beginning of actuarial profession. Actuaries use different mathematical models comparing the theoretical results towards the actual experience in order to intervene into the system and do some corrections. Generally, there are different types of actuarial investigations which are used as "control tools" (see Trowbridge (1989) i.e.

(i) "Participating Insurance": In the pricing procedure of a long term individual policy, an actuary should consider a conservative technical basis (low interest rates, high mortality rates high expense margins) for the need of premium calculations. At the valuation dates, he examines whether there is a gain which may be distributed as a "dividend" to the policyholder.

(ii) "Experience Rating": In group business an opposite approach is adopted using a short-term guarantee for the initial premium. Then on the renewals dates, negotiations take place with the policy holder in order to obtain an agreement for "a fair premium" (which heavily rely on recent claim experience).

(iii) "Actuarial gain/loss adjustment": This is a typical method used by pension actuaries in order to "handle" a pension plan. At each valuation date, actuaries calculate the surplus (overfunded) or deficit (underfunded) of a pension fund and action is decided for the future contribution rates.

We should also mention the bonus-malus system which is used (mostly) in motor-insurance as an other premium control procedure.

In our problem a possible control action should be similar to the action described as "experience rating".

Assuming we have a more "complete" information about the distribution of total annual claims (i.e. $C_n \sim N(\mu, \sigma^2)$) we may use Credibility theory in order to achieve a better estimation. As a general argument we may say that Credibility theory weights the actual experience towards an initial estimate.

At this point we shall provide the Bayesian approach to credibility theory quoting our results from Waters (1986).

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a normal distribution with density $f(x, \mu, \sigma^2)$. Then starting with a priori distribution for μ say $N(\mu_0, \sigma_0^2)$ we may find the posteriori distribution.

Finally we may obtain the mean of the posteriori distribution as the best estimation for the parameter μ and consequently for the required premium's estimation.

Hence,
$$\hat{C}_{n+1} = \left(\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \Bigg/ \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \tag{4.6.1}$$

and using (4.4.2) we obtain P_{n+1}.

So, actually, there are some "primitive" control techniques using standard decision rules which they are applied every n (n=1,2,3,...) year from the insurance managers trying to achieve the desired smooth results. A "manual" controller (the insurance manager) is attached into the system to protect it from undesired dangerous deviations.

The insurance manager always uses some standard strict mathematical criteria deciding for the right action on the right time. Of course a big question stands in front of us.

"How do we know that he decides the right action?" ... or another more difficult question.

"How do we know that the whole sequence of his right actions (decisions of each year) is the right sequence?"

It is a complicated matter of the choice of the "best strategy".The next section will give us another view of this problem.

4.7 An alternative approach based on utility theory or economics of insurance

In order to complete the discussion for the different methods of pricing of an insurance product, we should also mention two other elegant theories.

(i) Utility Theory: (Bowers et al (1986)). The basic assumption of this theory is the existence of a utility function say u which matches the wealth of amount w (measured in money units) to a numerical value (in most times) different from w and this function is unique for each individual (decision maker). Then the maximum premium P which is prepared to pay the individual above with wealth w in order to be totally covered from a random loss C with a distribution function H is determined by the equation:

$$u(w-P)=E(u(w-C)) \tag{4.7.1}$$

or $$u(w-P)=\int_{-\infty}^{+\infty} u(w-c)\,dH(c) \tag{4.7.2}$$

(ii) Economics of Insurance: Another interesting view is to consider the insurance product as a commodity which obeys the laws of the market with respect to its price. i.e. The price is determined by demand and supply functions (at the intersection point). Of course under this approach we may ignore everything else and just examine the market movements in order to decide our premium for the next year (hopefully near the mean).

4.8 The great challenge!... "Control the controller!"

A Greek (very popular nowadays) word may describe and give us a first explanation to the title of this section. The word is ..."Cybernetics"! But what do we mean?... and what is the great challenge?

For the first time, we can develop a theoretical model for the cybernetics of an insurance system. That means we shall be able to decide about the control strategy. Is it a good one? or is it a safe one? or ideally is it the best? (according to a certain set of criteria). For the first time we are able to control the controller.

Avoiding thousands of simulations we may reach directly the answer of the best strategy! Simulation procedures are time consuming and do not give us 100% safety about our decision making procedure.

An automatic controller (like an automatic pilot) is attached to the system and becomes responsible for the cybernetics of it. Although these automatic controllers may be very powerful and clever they can not substitute entirely the human management of the insurance company. So what is left for the managers?

The answer is very simple! Think of a modern airplane. They all have a human and an automatic pilot. Each of them takes action the right time and of course the human remains responsible for everything, especially during dangerous phenomena like a storm or a snowfall or...

Conclusively the role of an insurance manager is twofold.

(i) Develop, test and put in action the automatic controller in the good "sunny" days and

(ii) Put out of order the automatic controller and guide the system with his own arms through a terrible market "storm or snowfall".

4.9 Analysis of the model using control theory techniques

We shall consider the general version of our problem as described in section (4.3) with the restriction of a constant interest factor R without fluctuations adding also another four assumptions.

(i) No distribution of surplus to policyholders.

(ii) An inflation factor (F) for claim development (where F=1+inflation rate).

(iii) There is a "binding agreement" between the insurance company and the policy holders stating that all insurance contracts will remain in force up to infinity. This assumption is very important because it prevents withdrawal of the portfolio when the premium is risen due to the feedback of a negative surplus reserve.

(iv) There is a time delay factor (f) in the system. That means, we obtain the full information for the unknown variables of the system (i.e. claims with a time delay. So, at each time point n we have exact information up to time point n-f.

The last assumption is very important and may not actually hold in practice but it must be assumed here. In the real world, it is impossible to avoid the time delay element for a number of reasons which are well described in Ackman et al (1985) and be repeated below:

(i) "The insured contingency itself may not occur at a single instant - for example workmen's compensation".

(ii) "The legal liability of the insurer may not always be clear cut and there may be considerable delays before the insurer (or the court) decides that a liability exists"

(iii) "The quantum of damages may be impossible to determine until some time has elapsed since the occurrence of the event - for example motor damage claims..."

(iv) "There will be processing delays within the insurer's office..."

(v) "There may be delays before the insurer is even notified that a claimable event has occurred..."

Firstly, we should establish a standard decision rule in order to produce a first estimation for the next year's risk premium $\hat{C}^{1}_{n+1}$. Although we can use complicated mathematical rules we prefer a relative simple one described as the weighted average prognosis based on the last p years of actual claim experience. The weights shall have the format described in equation (4.4.4) and will be applied to claims experience $C_{n-p-f+1}$, $C_{n-p-f+2}, \ldots, C_{n-f}$ as we have introduced the time delay i.e.

$$\hat{C}^{1}_{n+1} = \frac{1}{M}\left[F^{p+f} \cdot \theta \cdot C_{n-p-f+1} + F^{p+f-1}\theta^{2} \cdot C_{n-p-f+2} + \ldots + F^{1+f} \cdot \theta^{p} \cdot C_{n-f}\right] \qquad (4.9.1)$$

where $$M = \sum_{k=1}^{p} F^{p+f+1-k}\theta^{k}$$

and subsequently we may produce another estimate taking into account the expense margin $\hat{C}^2_{n+1}$ i.e.

$$\hat{C}^2_{n+1} = \frac{1}{e} \cdot \hat{C}^1_{n+1} \tag{4.9.2}$$

Our basic aim is to design such a control mechanism in order to obtain the ideal smoothness for both premiums and surplus. Finally, we prefer a stable system which shall be self-protected from instability. This may be obtained inserting a negative feed-back mechanism of the accumulated surplus. Before go further with the concept of feedback, we shall do some basic control analysis of the problem.

In our example the amount of total annual claims is considered to be the input of the system. The process is all the set of equations, functions and logical rules which produce the output of the system. The output consists of two variables: premium and surplus. As it usually happens, the smoothness of the first output variable (surplus) operates contrariwise to the smoothness of the second output variable (premiums).

A measuring procedure on the surplus output will enable us to decide whether our premium estimates are correct or not. This information of the output should be fed back. But the big questions are: How? and How much of this information should be fed back in order to stabilize the system? We shall try below to answer these questions.

The annual surplus L_n is given by the formula

$$L_{n+1}=eRP_{n+1}-C_{n+1} \tag{4.9.3}$$

and the accumulated surplus S_{n+1} at the end of year (n+1)

$$S_{n+1}=RS_n+L_{n+1} \tag{4.9.4}$$

Finally, we may refund a fraction (say ε) of this accumulated surplus to the policyholder.

Hence, $$P_{n+1} = \hat{C}^2_{n+1} - \varepsilon S_{n-f} \tag{4.9.5}$$

where $|\varepsilon|$ normally lies in the interval [0,1]. Again we identify the existence of the delay factor (f) for the accumulated surplus (S). (We are at the end of year n, without knowing the actual claim experience for years n, n-1,...,n-f+1 consequently the last known value for the accumulated surplus is the S_{n-f}).

As we have stated the current model is an extension of the one developed by Balzer & Benjamin (1980) generalized in three directions.

(1) There is an additional parameter R which reflects the investment income falling into the system.

(2) The "exponential smoothing" process is used (with F,θ inflation & weighting factors) as an advanced claim predictor providing smoother estimations.

(3) The delay (f) is a free control parameter which value will be examined in conjunction with the stability of the system.

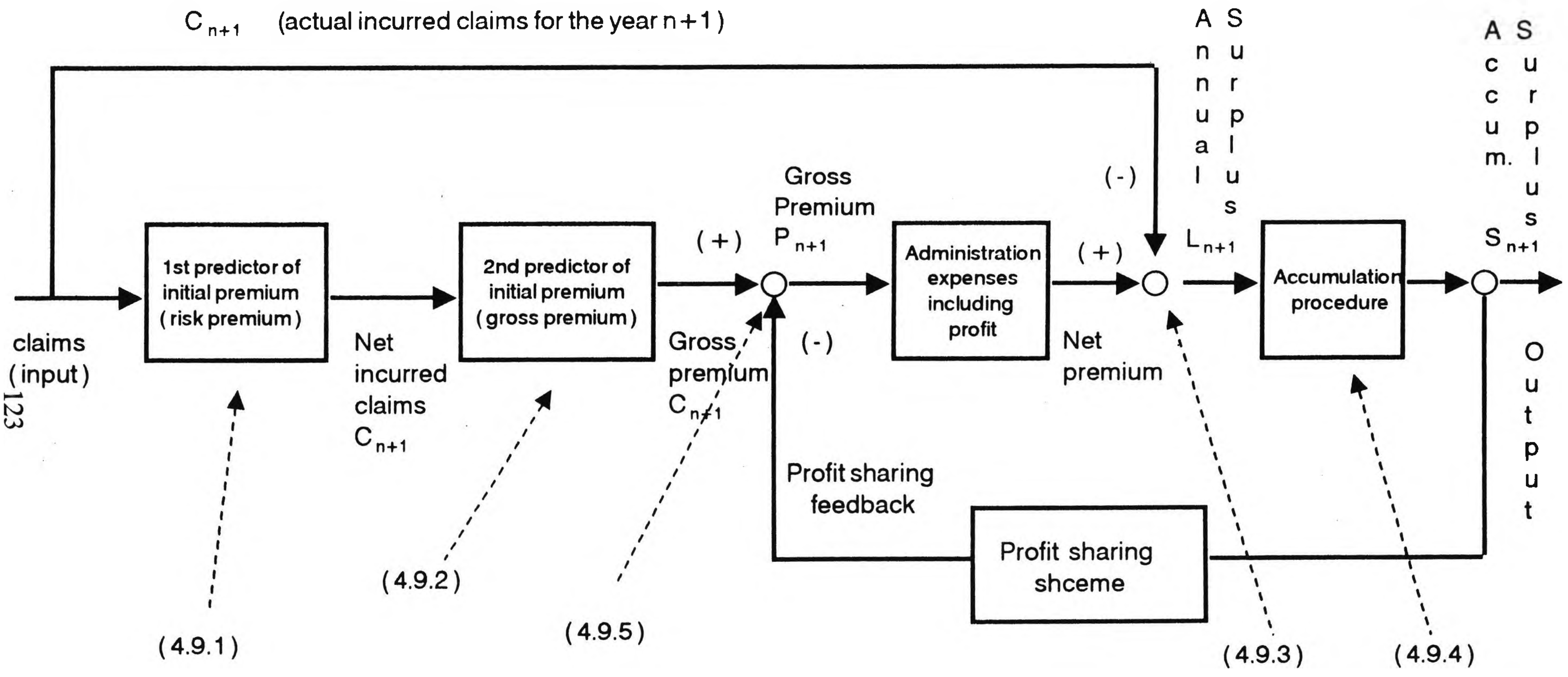
4.10 Control diagram of the model

Now, we shall draw the control diagram of the model in order to obtain a geometrical representation of the process (see Diagram (4.10.1) of the next page).

We may picture another representation of the diagram by putting into the boxes the algebraic equations and indicating the arrows with the analogous variable.

Finally, we may have the formal representation of a control diagram, which shows the z-transformation of each equation into the boxes.

Control diagram of the model



In the Diagram we have indicated with dotted arrows the relationship between boxes or circles with the equations of section (4.9)

Diagram (4.10.1)

4.11 The general solution to the system of equations

Considering all the equations which have been described in section (4.9) and changing slightly the subscripts, we may produce the following system of equations.

$$\hat{C}_n^1 = \frac{1}{M}\left[F^{p+f}\theta \cdot C_{n-p-f} + F^{p+f-1}\theta^2 \cdot C_{n-p-f+1} + \ldots F^{1+f}\theta^p \cdot C_{n-f-1}\right] \qquad (4.11.1)$$

$$\hat{C}_n^2 = \frac{1}{e}\hat{C}_n^1 \qquad (4.11.2)$$

$$P_n = \hat{C}_n^2 - \varepsilon S_{n-f-1} \qquad (4.11.3)$$

$$L_n = eRP_n - C_n \qquad (4.11.4)$$

$$S_n = RS_{n-1} + L_n \qquad (4.11.5)$$

Now combining equations (4.11.1), (4.11.2) and (4.11.3) we obtain,

$$P_n = \frac{1}{M \cdot e}\left[F^{p+f}\theta \cdot C_{n-p-f} + \ldots + F^{1+f}\theta^p \cdot C_{n-f-1}\right] - \varepsilon S_{n-f-1} \qquad (4.11.6)$$

and combining equations (4.11.4) & (4.11.5) we obtain

$$S_n = RS_{n-1} + eRP_n - C_n \qquad (4.11.7)$$

finally from equations (4.11.6) & (4.11.7) we arrive at

$$S_n = RS_{n-1} + \frac{R}{M}\left[F^{p+f}\theta \cdot C_{n-p-f} + \ldots + F^{1+f}\theta^p \cdot C_{n-f-1}\right] - eR\varepsilon S_{n-f-1} - C_n \qquad (4.11.8)$$

The last expression is linear time-invariant difference equation. We shall proceed to its solution by considering the z-transformation which will be defined here slightly differently as

$$S_z = \mathscr{Z}\{S_n\} = \sum_{n=0}^{\infty} S_n z^n \text{ , when } |z| < 1$$

(similarly is defined the C_z).

Now assuming zero initials conditions i.e.

$C_{-k}=0$ and $S_{-k}=0$ for every k=0,1,2,...

we may obtain the following equation for the z-transformed variables S_z and C_z i.e.

$$S_z = RzS_z + \frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{l+f}\theta^p z^{l+f}\right]C_z - eR\varepsilon z^{l+f}S_z - C_z \tag{4.11.9}$$

the last equation is rearranged putting the terms of S_z into the left-hand side while the terms of C_z into the other side. So,

$$\left(1 - Rz + eR\varepsilon z^{l+f}\right)\cdot S_z = \left[\frac{R}{M}\left(F^{p+f}\theta z^{p+f} + \ldots + F^{l+f}\theta^p z^{l+f}\right) - 1\right]\cdot C_z \tag{4.11.10}$$

or equivalently,

$$S_z = -\frac{1 - \frac{R}{M}\left(F^{p+f}\theta z^{p+f} + \ldots + F^{l+f}\theta^p z^{l+f}\right)}{1 - Rz + eR\varepsilon z^{l+f}}C_z \tag{4.11.11}$$

The fraction (with the negative sign) which is included in the last expression (4.11.11) is the transfer function of the system

(transfer function): $$G(z) = -\frac{1 - \frac{R}{M}\left(F^{p+f}\theta z^{p+f} + \ldots + F^{l+f}\theta^p z^{l+f}\right)}{1 - Rz + eR\varepsilon z^{l+f}} \tag{4.11.12}$$

As we have seen in chapter 2, G(z) is very important for the investigation of the system as represents the basic cybernetic element containing the full information of the whole process.

A typical control diagram using the G(z) function is given below:

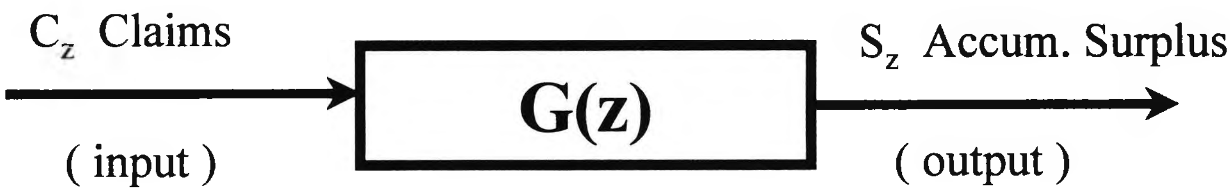


Diagram (4.11.1)

In order to proceed with the solution of the initial equation (4.11.8) we have to consider now the roots of the characteristic polynomial (or equivalently the poles of the transfer function). Now let $\frac{1}{\rho_1}, \frac{1}{\rho_2}, \ldots, \frac{1}{\rho_n}$ be the roots of equation

$$h(z) = 1 - Rz + eR\varepsilon z^{1+f} = 0 \tag{4.11.13}$$

then it is easily proved that $\rho_1, \rho_2, \ldots, \rho_n$ are the roots of

$$\varphi(z) = z^{1+f} - Rz^{f} + eR\varepsilon = 0 \tag{4.11.14}$$

and we may split the transfer function G(z) by the methods of partial fractions and re-write (4.11.11) in the following general form (see Goldberg (1963)).

$$\begin{aligned} S_z = & \left[\frac{a_{11}}{1-\rho_1 z} + \frac{a_{12}}{1-\rho_2 z} + \ldots + \frac{a_{1,1+f}}{1-\rho_{1+f} z} \right] C_z + \\ & + \left[\frac{a_{21}}{1-\rho_1 z} + \frac{a_{22}}{1-\rho_2 z} + \ldots + \frac{a_{2,1+f}}{1-\rho_{1+f} z} \right] z^{1+f} C_z + \ldots \\ & \ldots \qquad \ldots \qquad \ldots \\ & + \left[\frac{a_{p+1,1}}{1-\rho_1 z} + \frac{a_{p+1,2}}{1-\rho_2 z} + \ldots + \frac{a_{p+1,1+f}}{1-\rho_{1+f} z} \right] z^{p+f} C_z \end{aligned} \tag{4.11.15}$$

Finally, we may obtain the solution by inverting the z-transformation of last equation (4.11.15) and constructing the general form of S_n (see Goldberg (1963)) as,

$$
\begin{aligned}
S_n = & \left(a_{11}\rho_1^n + a_{12}\rho_2^n + \ldots + a_{1,1+f}\rho_{1+f}^n\right) * C_n + \\
& + \left(a_{21}\rho_1^n + a_{22}\rho_2^n + \ldots + a_{2,1+f}\rho_{1+f}^n\right) * C_{n-f-1} + \ldots \\
& \ldots \qquad \ldots \qquad \ldots \\
& + \left(a_{p+1,1}\rho_1^n + a_{p+1,2}\rho_2^n + \ldots + a_{p+1,1+f}\rho_{1+f}^n\right) * C_{n-f-p}
\end{aligned} \tag{4.11.16}
$$

where

- coefficients a_{ij} $i=1,2,\ldots,p+1$, $j=1,2,\ldots,1+f$ are determined by the method of partial fractions under the following relationship.

  The values of a_{1j} $j=1,2,\ldots,1+f$ are calculated, equating the expression which contains them with -1 while

  for each $i=2,3,\ldots,p+1$ the values of a_{ij} $j=1,\ldots,1+f$ are calculated, equating the expression which contains them with $\frac{R}{M}\cdot F^{i+f-1}$

  i.e. a_{ij} analogous to $\frac{R}{M}\cdot F^{i+f-1}$ (4.11.17)

- and the sign * which appears in equation (4.11.16) means convolution!

4.12 Comments on the general solution

In this section and after having obtained the solution of the problem, we shall make short comments to the general format and on each parameter involved with its possible affect.

Solution structure: Generally speaking, the solution depends on three factors.

(a) Past Claim experience (linearly)

(b) The roots of equation (4.11.14) (exponentially)

(c) The coefficients a_{ij} $i=1,\ldots,1+f$, $j=1,\ldots,p+1$ (linearly)

As we can see the most important factor is the second one, the roots of the characteristic equation which has an exponential affect. Now since we can not alter the (a) factor we shall study the (b), (c) factors in conjunction with the parameters of the problem: $f, R, e, \varepsilon, F, p, \theta$.

f delay factor: The delay factor appears to be the most important feature of the whole process as affects the degree of the characteristic polynomial (consequently the number of roots of characteristic equation). The effect is exponential because the roots appear in the power format in the solution. Another interesting result is the appearance of certain oscillations if $f \geq 3$. It has been shown in Zimbidis & Haberman (1993) that equation (4.11.14) has at least one pair of conjugate complex number (using Descartes Rule of Signs - positive roots). So, if $f \geq 3$ then the general solution (i.e. the accumulated surplus and consequently the premiums) will exhibit some kind of oscillations (longer discussion will follow in later section using the root locus method).

R interest factor: The interest factor appears both in the characteristic equation and in the expression which determines the coefficients a_{ij} so it has a multiple effect in the solution. As R appears in the numerator of the right hand side of expression (4.11.17) its effect should be determined analogously. Generally speaking R accelerates the surplus procedure i.e. Larger values of R result in greater output responses (and similarly for smaller values). The effect of R into the characteristic equation will be discussed in later section with the help of the root locus method. R could also be a function of time but this problem will be dealt with in sections (4.20) & (4.21).

e,ε expense and feedback factors: These appear in the characteristic equation so have an exponential effect as they determine the position of the roots (full discussion in a later section using the root locus method).

θ, F weighting and inflation factors: They appear in expression (4.11.17) both in the numerator and the denominator of the respective fraction. They affect the coefficients a_{ij}, so actually determine the amplitude of the output (S_n). Large values of θ,F will produce large coefficients a_{ij} when i's are small, so actually the output will be affected from recent claim experience (expected also by general reasoning)

p averaging factor: It appears in many places of the solution determining the "length" of equation (4.11.16). Large values of p produce many coefficients a_{ij} consequently more claim experience is incorporated in the solution and finally smoother results for the output of the system.

Control of parameters: Having examined the possible effects of each parameter the obvious question is whether we can control them and in a way "guide the solution". Generally, we may separate the parameters above in three categories.

1st category: (R,e,F) These parameters may be almost uncontrollable (or controlled but in a very thin interval) by the actuary.

2nd category: (f,p,θ) These parameters may be controllable especially when there is enough past data, so an actuary may choose the delay factor or the averaging number of years, or the special weighting factor θ.

3rd category: (ε) The feedback factor may be determined and be absolutely controlled by the actuary.

After having established the three categories above the actuary (the person who creates and follows the system) may design the optimal control strategy in order to obtain the desired results for surplus and premium variables (i.e. smooth, stable etc.)

Closing this section we must distinguish the two meanings of a controllable parameter (the potential to determine its numerical value without any restrictions) and controllable system which has been defined in section (2.13)

4.13 Analysis of a special case (p=2, f=1, θ=1)

Before go further with more control analysis of the general solution let's try to have a first "look" (in numerical results) using simulation techniques and a special case of the model where p=2, f=1 & θ=1.

Consequently the system of equations (4.11.1) - (4.11.5) will be realized as follows:

$$\hat{C}_n^1 = \frac{1}{M'}\left[F^3 C_{n-3} + F^2 C_{n-2}\right] \tag{4.13.1}$$

$$\hat{C}_n^2 = \frac{1}{e}\hat{C}_n^1 \tag{4.13.2}$$

$$P_n = \hat{C}_n^2 - \varepsilon S_{n-2} \tag{4.13.3}$$

$$L_n = eRP_n - C_n \tag{4.13.4}$$

$$S_n = RS_{n-1} + L_n \tag{4.13.5}$$

where $M' = F^2 + F^3$

We may assume that the system stands on the equilibrium point i.e.

$$C_n=0\,, \quad P_n=0\,, \quad S_n=0\,, \quad \forall n\leq 0$$

and simulate it, using the recursive relationship of equation (4.11.8) (replacing p=2 θ=1 and f=1, so obtain (4.13.6)) under different claim patterns i.e.

$$S_n = RS_{n-1} - eR\varepsilon S_{n-2} + \frac{R}{M'}\left(F^3C_{n-3} + F^2C_{n-2}\right) - C_n \tag{4.13.6}$$

We may produce tables (and diagrams respectively using different values for the set of the basic parameters $\{R,e,F,\varepsilon\}$. We shall use some typical values (for Greece in years 1990-92, or generally for an economy with high inflation and interest rates and with insurance companies having a substantial expense & profit margin) i.e.

$$1.18 \leq R \leq 1.22, \quad 60\% \leq e \leq 80\%, \quad 1.12 \leq F \leq 1.15$$

For the feedback function, we shall consider values from 10% up to 100%, while the following patterns for claims will be considered:

(i) a spike function where

$C_0=0, C_1=1, C_2=0, C_3=0, \ldots, C_n=0, \ldots$

(ii) a step function where

$C_0=0, C_1=1, C_2=1, C_3=1, \ldots, C_n=1, \ldots$

(iii) a random variable uniformly distributed in the unit interval i.e.

$C_n \sim U(0,1) \qquad n=0,1,2,\ldots$

A computer function will be used to generate the observations of the uniform distribution.

In the next section we present the results and the respective diagrams of the simulations.

4.14 Presentation of some examples with brief comments (Tables & Diagrams)

1st example: (input signal = spike)

The first example contains a spike function as an input signal. The three other parameters R,e,F are constant while we examine four different values for the parameter ε.

Hence, we simulate four different set of values as the following table (4.14.1)

Table (4.14.1)

	Input Signal	R	e	F	ε
(i)	Spike	1.18	60%	1.12	10%
(ii)	Spike	« «	« «	« «	30%
(iii)	Spike	« «	« «	« «	50%
(iv)	Spike	« «	« «	« «	100%

The results of our simulations are shown in table (4.14.2). There are six columns. The first one represents the development of the time variable. The second one, shows the claim amount at each time t. Finally the last four columns which correspond to the rows (i), (ii), (iii), (iv) of the table (4.14.1.) show the development of the surplus variable S_n, according to the selected value of ε=0.1, 0.3, 0.5 or 1 (i.e. 10%, 30%, 50%, 100%).

We shall also provide table (4.14.3.) which shows the development of the two variables C_n and S_n for general values of the parameters R,e,F,ε and a spike input signal.

TABLE (4.14.2)

Input Signal ***Spike*** ***(0,1,0,0,0,0,...)***

Expense	***(e)***	***60%***
Interest	***(R)***	***1,18***
Inflation	***(F)***	***1,12***
	(M/)	***2,659***

		(i)	(ii)	(iii)	(iv)
Feedback (ε)		**10%**	**30%**	**50%**	**100%**
Time	**Claims**	**Accumulated surplus**	**Accumulated surplus**	**Accumulated surplus**	**Accumulated surplus**
0	0	0,0000	0,0000	0,0000	0,0000
1	1	-1,0000	-1,0000	-1,0000	-1,0000
2	0	-1,1800	-1,1800	-1,1800	-1,1800
3	0	-0,7650	-0,6234	-0,4818	-0,1278
4	0	-0,1958	0,1384	0,4726	1,3080
5	0	-0,1768	0,2957	0,7282	1,6340
6	0	-0,1948	0,3196	0,6920	1,0020
7	0	-0,2173	0,3143	0,5588	0,0255
8	0	-0,2427	0,3030	0,4144	-0,6793
9	0	-0,2710	0,2908	0,2912	-0,8196
10	0	-0,3026	0,2787	0,1969	-0,4862
11	0	-0,3378	0,2672	0,1293	0,0066
12	0	-0,3772	0,2560	0,0828	0,3520
13	0	-0,4212	0,2454	0,0520	0,4107
14	0	-0,4703	0,2352	0,0320	0,2354
15	0	-0,5252	0,2254	0,0194	-0,0130
16	0	-0,5864	0,2160	0,0115	-0,1820
17	0	-0,6547	0,2070	0,0067	-0,2056
18	0	-0,7311	0,1984	0,0039	-0,1137
19	0	-0,8163	0,1901	0,0022	0,0114
20	0	-0,9115	0,1822	0,0012	0,0939
21	0	-1,0178	0,1746	0,0007	0,1028
22	0	-1,1364	0,1674	0,0003	0,0548
23	0	-1,2689	0,1604	0,0002	-0,0081
24	0	-1,4169	0,1537	0,0001	-0,0484
25	0	-1,5821	0,1473	0,0000	-0,0513
26	0	-1,7666	0,1412	0,0000	-0,0263
27	0	-1,9725	0,1353	0,0000	0,0053
28	0	-2,2025	0,1297	-0,0000	0,0249
29	0	-2,4593	0,1243	-0,0000	0,0256
30	0	-2,7460	0,1191	-0,0000	0,0126
31	0	-3,0662	0,1142	-0,0000	-0,0033
32	0	-3,4237	0,1094	-0,0000	-0,0128
33	0	-3,8229	0,1048	-0,0000	-0,0128
34	0	-4,2686	0,1005	-0,0000	-0,0060
35	0	-4,7663	0,0963	-0,0000	0,0019
36	0	-5,3220	0,0923	-0,0000	0,0065
37	0	-5,9425	0,0885	-0,0000	0,0063
38	0	-6,6354	0,0848	-0,0000	0,0029
39	0	-7,4090	0,0812	-0,0000	-0,0011
40	0	-8,2729	0,0779	-0,0000	-0,0033

Table (4.14.3)

time (n)	claims (C_n)	Accumulated Surplus variable S_n
0	0	0
1	1	-1
2	0	-R
3	0	$-R^2 + R \cdot \left(e \cdot \varepsilon + \frac{F^2}{M'} \right)$
⋮	⋮	⋮

We have also drawn the diagrams (4.14.1) and (4.14.2) which show the graphical representation of columns (i)-(iv) of table (4.14.2). Consequently we have the following comments.

(1) As we observe in diagram (4.14.1) the surplus variable goes to minus infinity. That means, if we use a small percentage (e.g. ε=10%) as feedback the system will never reach the stable state (action: avoid small values of ε).

(2) As we observe in diagram (4.14.2) the surplus variable converges to zero for every value of ε=30%, 50%, 100%. But the process shows oscillatory behavior for the large value of ε=100% (action: avoid large values of ε).

(3) Obviously, a selection between 30% and 50% for the feedback parameter will be the best choice (of the ones examined). If we would like to choose just one value, we should consider other features of the system. For instance, we may consider the settlement time (how quickly something goes to zero) for each value of ε. In our graph we observe that the path of 50% feedback goes faster.

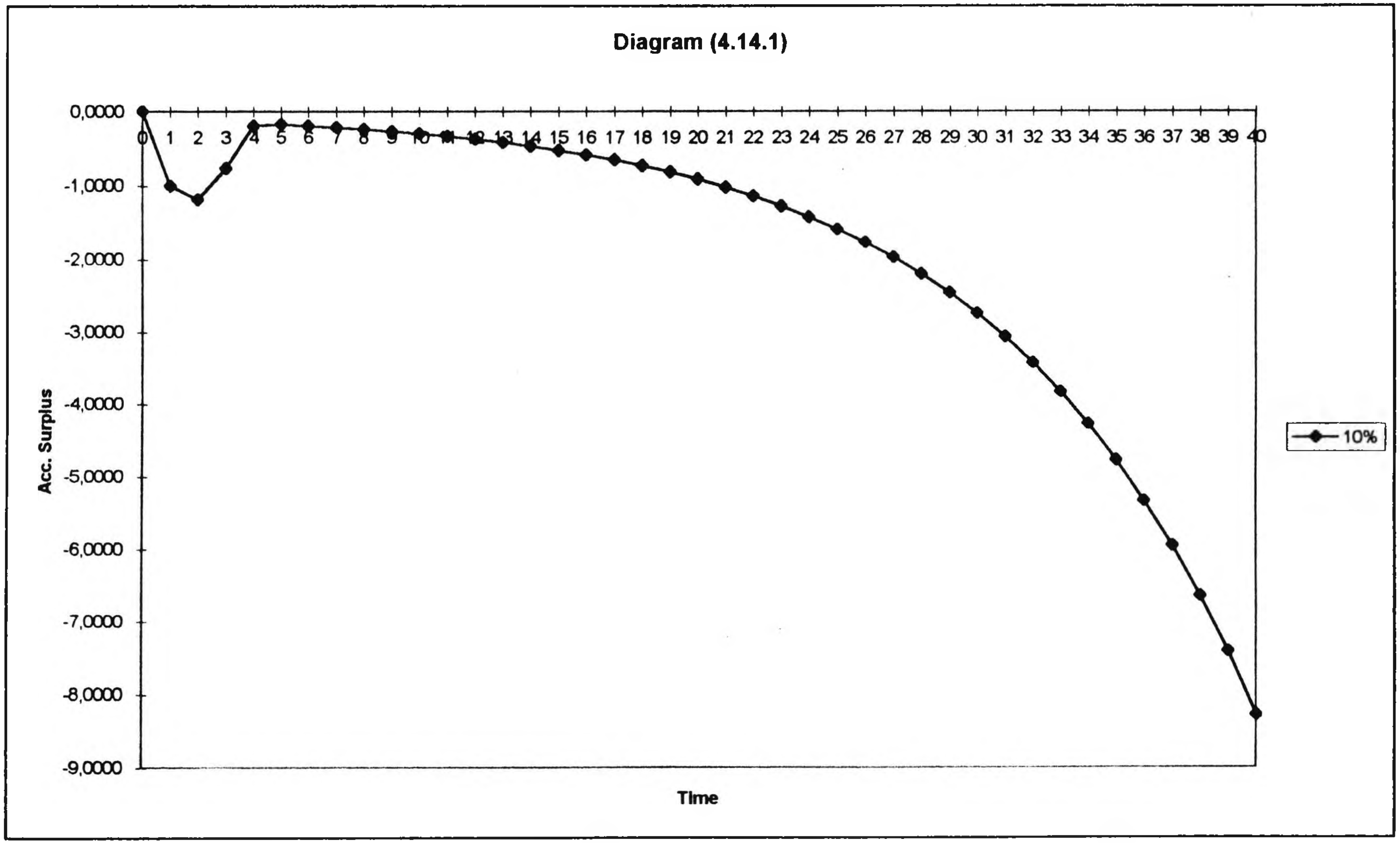


Diagram (4.14.1)
Acc. Surplus
0,0000
-1,0000
-2,0000
-3,0000
-4,0000
-5,0000
-6,0000
-7,0000
-8,0000
-9,0000
Time
10%

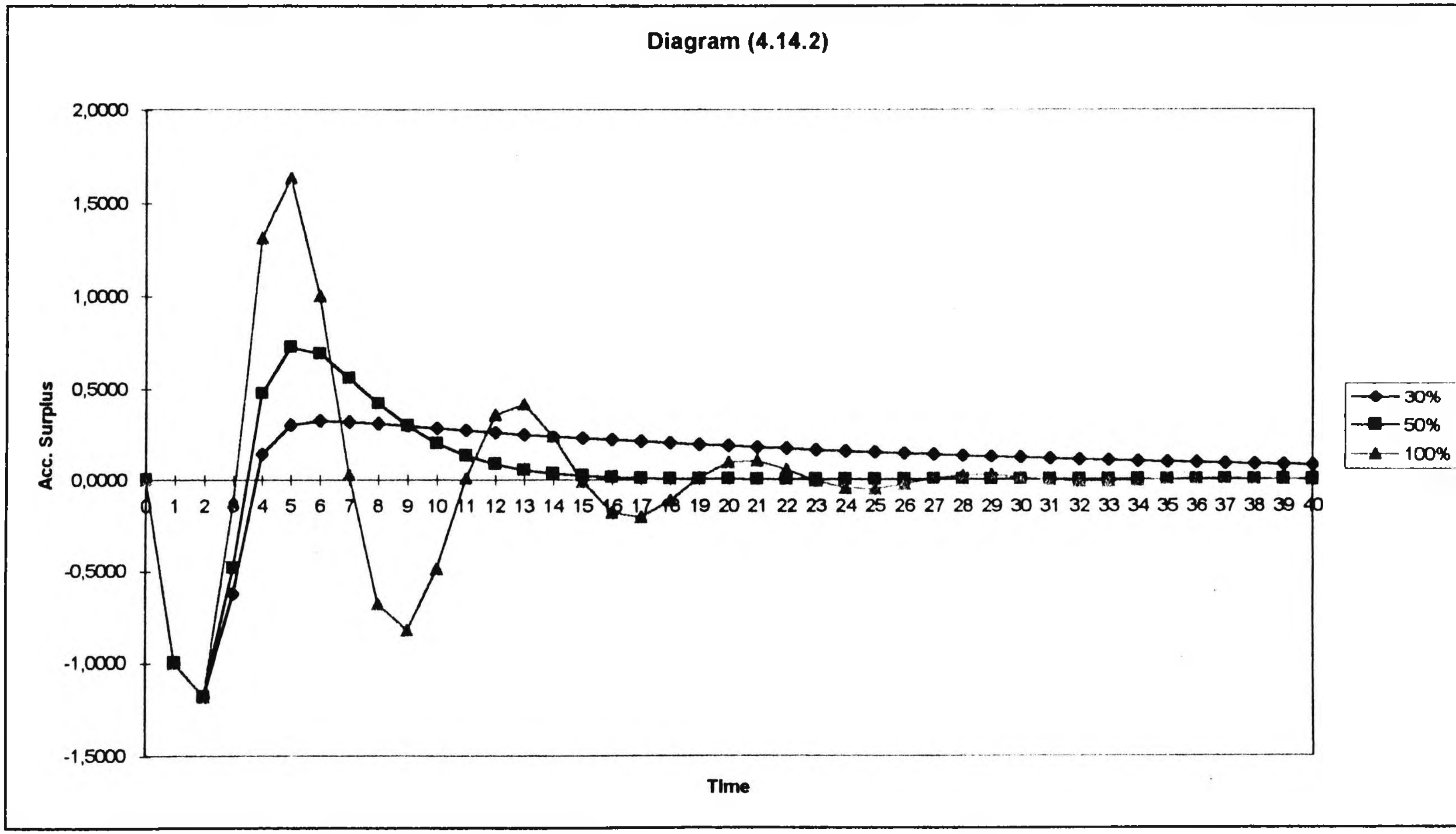


Diagram (4.14.2)
2,0000
1,5000
1,0000
0,5000
0,0000
-0,5000
-1,0000
-1,5000
Acc. Surplus
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Time
30%
50%
100%

2^{nd} example: (input signal = step)

The second example contains a step function as an input signal. Similarly with the 1^{st} simulation obtain again four different set of values as the following table (4.14.4)

Table (4.14.4)

	Input Signal	R	e	F	ε
(i)	step	1.22	80%	1.15	30%
(ii)	« «	« «	« «	« «	50%
(iii)	« «	« «	« «	« «	80%
(iv)	« «	« «	« «	« «	100%

The results of our simulations are shown in table (4.14.5). (The format of the table (4.14.5) is exactly the same as for table (4.14.2)).

We shall also provide table (4.14.6) similar to (4.14.3).

Table (4.14.6)

time (n)	claims (Cn)	Accumulated Surplus variable S_n
0	0	0
1	1	-1
2	1	-(1+R)
3	1	$-R^2+R\cdot\left(e\cdot\varepsilon+\frac{F^2}{M'}-1\right)-1$
⋮	⋮	⋮

We also present diagram (4.14.3) including the graphical representation of table (4.14.5).

TABLE (4.14.5)

Input Signal ***Step*** ***(0,1,1,1,1,1,...)***

Expense	***(e)***	***80%***
Interest	***(R)***	***1,22***
Inflation	***(F)***	***1,15***
	(M/)	***2,843***

Feedback (ε)		***(I)*** ***30%***	***(ii)*** ***50%***	***(iii)*** ***80%***	***(iv)*** ***100%***
Time	***Claims***	***Accumulated surplus***	***Accumulated surplus***	***Accumulated surplus***	***Accumulated surplus***
0	0	0,0000	0,0000	0,0000	0,0000
1	1	-1,0000	-1,0000	-1,0000	-1,0000
2	1	-2,2200	-2,2200	-2,2200	-2,2200
3	1	-2,8482	-2,6530	-2,3602	-2,1650
4	1	-2,6047	-1,9332	-0,9260	-0,2545
5	1	-2,1238	-0,8439	0,9331	2,0225
6	1	-1,6084	0,1338	2,0814	2,9358
7	1	-1,1204	0,7951	2,0307	1,8278
8	1	-0,6760	1,1247	1,0724	-0,4155
9	1	-0,2766	1,2042	-0,0573	-2,0708
10	1	0,0805	1,1402	-0,6872	-1,9009
11	1	0,3991	1,0234	-0,5737	-0,0780
12	1	0,6834	0,9121	0,0567	1,9801
13	1	0,9369	0,8334	0,7371	2,7119
14	1	1,1629	0,7916	1,0750	1,5959
15	1	1,3644	0,7791	0,9560	-0,4798
16	1	1,5441	0,7842	0,5469	-1,9230
17	1	1,7043	0,7965	0,1408	-1,6577
18	1	1,8471	0,8090	-0,0352	0,0744
19	1	1,9745	0,8183	0,0671	1,9287
20	1	2,0880	0,8236	0,3293	2,5004
21	1	2,1893	0,8254	0,5694	1,3880
22	1	2,2795	0,8251	0,6575	-0,5269
23	1	2,3600	0,8238	0,5776	-1,7776
24	1	2,4318	0,8224	0,4113	-1,4344
25	1	2,4957	0,8213	0,2708	0,2050
26	1	2,5528	0,8207	0,2292	1,8701
27	1	2,6036	0,8204	0,2882	2,3014
28	1	2,6490	0,8204	0,3927	1,2025
29	1	2,6894	0,8205	0,4740	-0,5591
30	1	2,7255	0,8207	0,4917	-1,6357
31	1	2,7576	0,8208	0,4498	-1,2299
32	1	2,7863	0,8209	0,3848	0,3160
33	1	2,8118	0,8209	0,3383	1,8059
34	1	2,8346	0,8209	0,3322	2,1148
35	1	2,8549	0,8209	0,3612	1,0375
36	1	2,8730	0,8209	0,4013	-0,5783
37	1	2,8892	0,8209	0,4275	-1,4981
38	1	2,9036	0,8209	0,4283	-1,0433
39	1	2,9164	0,8209	0,4087	0,4093
40	1	2,9278	0,8209	0,3842	1,7377

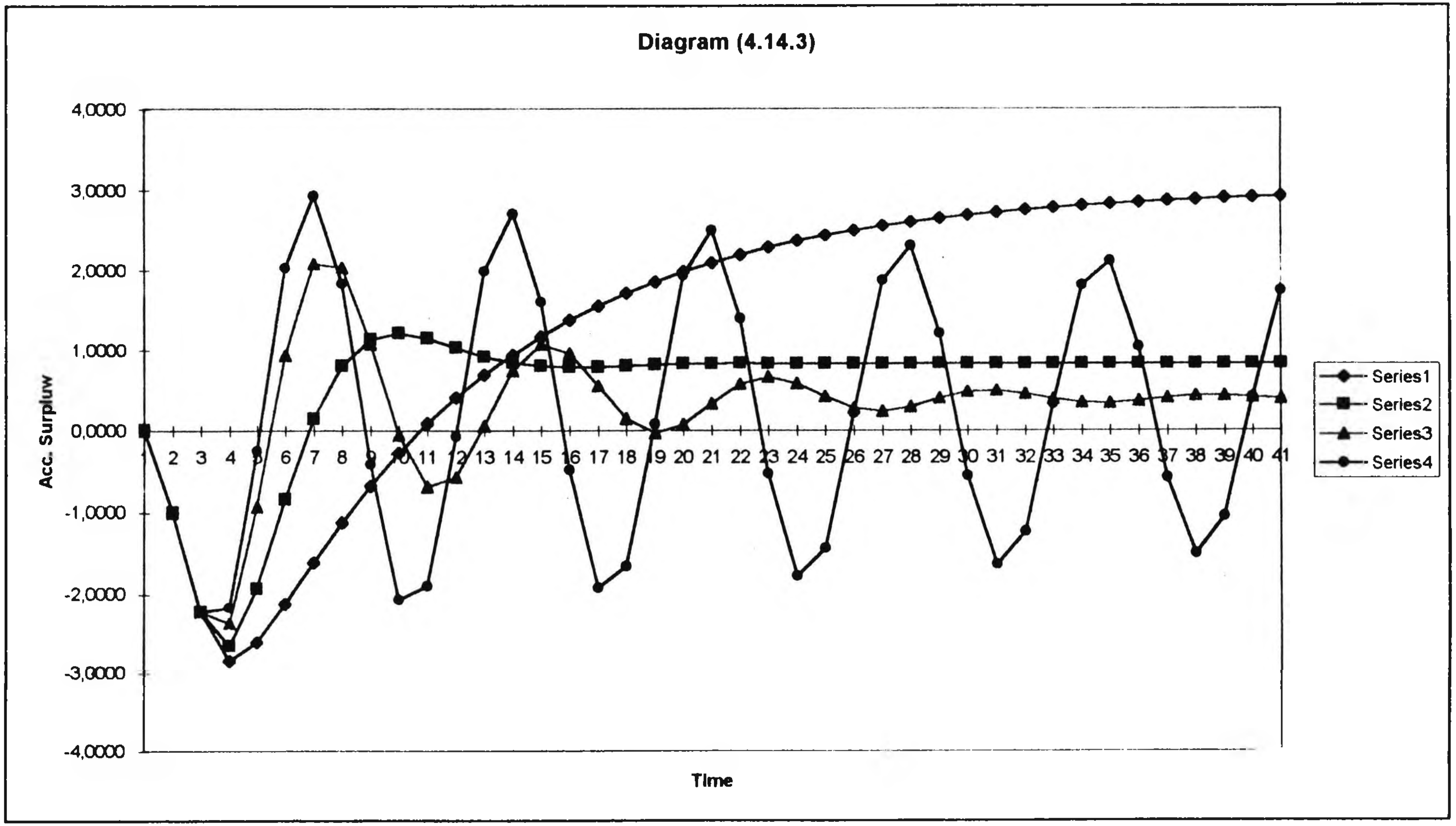


Diagram (4.14.3)

(1) As we observe in diagram (4.14.3), no one path converges to zero. That means the design of the system is not good enough to react properly to a step signal.

(2) Again, the path of 100% feedback (as for the spike input) exhibits oscillations (showing a slightly decreasing amplitude).

(3) The three paths of 30%, 50%, 80% feedback appear to converge to certain values. These limits will be examined in section (4.19). We shall find exactly how far from zero the path may converge.

(The limit as quoted in (4.19)): $\dfrac{\frac{R}{M}\left[F^{p+f}\theta + \ldots + F^{1+f}\cdot\theta^{p}\right] - 1}{1 - R + eR\varepsilon}$

As has been shown in Balzer (1982) for a similar model, integral action should be used in order to "bring the system back to initial condition of zero". In section (4.19) we also use integral action and we prove that system may return back to initial state of zero.

3^rd^ example: (input signal = random* U(0,1)*)

The third example contains a random (U(0,1)) function as an input signal.

We assume constant values for parameters R, e, F i.e.

$$R=1.20\ ,\quad e=70\%\ ,\quad F=1.14$$

and we run 1.000 simulations for ten different values of ε (feedback factor)

$$\varepsilon=10\%\ ,\ 20\%\ ,\ 30\%\ ,\ \ldots\ ,\ 90\%\ ,\ 100\%$$

Then for each simulation we find the value of ε (feedback factor) which produces the output (surplus) sequence of values with the minimum mean and minimum variance.

The results of our simulations with respect to the minimum mean and variance are shown in the following table (4.14.8).

(Interpreting table (4.14.8) we can say that under 81 simulations the output which corresponds to ε=20% had the minimum mean etc...).

Table (4.14.8)

MINIMUM MEAN			MINIMUM VARIANCE		
(ε) feedback factor	Simulations	% frequency	(ε) feedback factor	Simulations	% frequency
0,2	81	8,10%	0,4	38	3,80%
0,7	1	0,10%	0,5	484	48,40%
0,8	22	2,20%	0,6	305	30,50%
0,9	116	11,60%	0,7	107	10,70%
1	780	78,00%	0,8	44	4,40%
Total	1000		0,9	11	1,10%
			1	11	1,10%
			Total	1000	

As we can see the greater the value of (ε) then the greater probability to obtain the minimum value for the mean of the accumulated surplus. For example if ε=100% then in 780 cases (i.e. 78%) the minimum value for the mean is obtained.

As regard the variance of the accumulated surplus we observe that it is obtained (with greater probability) in the area of 50% and 60% of the (ε) feedback factor (the probability for the minimum variance under ε=50% or 60% is 78.9%).

Finally from our simulations we found that

(1) No path (any choice of e) appears to converge to a certain value.

(2) The path of 100% feedback has the most oscillatory form

4.15 The "Damokleios Sword"! (the great threat)

The simulation tables sometimes cause more confusion than solving the problem. Of course, we obtain a first "taste" but we have no insight to the behaviour of the system on a long-term basis. On the other hand we never feel absolutely safe (even in the short-term) because we can not simulate all the different sets of values for the parameters of the problem (e.g. in our model e,R,ε,F).

A "Damokleios sword" is always above our heads threatens with a catastrophe of the system. A "dangerous" set of values which has never been simulated may occur (e.g. F=1.35 high inflation rate) and destroy the whole process.

But even if we could avoid the total catastrophe we would not be happy again because we aim to be in a position to obtain the best strategy for the future designing the values of parameters which are partially or fully controlled by us.

As we have seen in section (4.12) ε feedback factor may be a fully controllable parameter. We should be able to choose the "best" ε for our system, considering the typical values of the other parameters.

The expense e factor is also a controllable factor, since it contains the desirable profit margin. An insurance company may modify (each year) its profit margin in order to obtain a better behaviour (and the desirable stability) of the system.

Of course, one may argue that in a very competitive market these factors may not be fully controllable. The competition may restrict the choices of ε and e.

Anyway, (partially or fully) there are some parameters which may be controlled, so we need a more advanced analysis of the system revealing the role of each parameter on a long-term basis. We shall then be in a position to design the parameters of the system and obtain the "best control strategy".

4.16 Stability analysis of the general model using the root locus method

The analytical solution (described by equation (4.11.16)) of the difference equation (4.11.8) may provide us full information about the behaviour of the system but in most times it is very difficult to obtain the exact numerical form. Fortunately, the most times it is needless because we do not need such full information. We only need to be able to answer some basic questions. For example the most important one: "Is the system stable?".

A theory has been developed which examines the qualities of difference (and differential) equations without solving them. In this section, we shall discuss the general solution of our system with respect to stability using the root locus method (Appendix III).

From equation (4.11.16) we may conclude that the behavior of the system depends (heavily) upon the roots of equation (4.11.14). (The dependency should be considered on a long-term basis). Firstly, there are three options for convergence:

(i) The absolute values of the roots are less than the unity i.e.

$$|\rho_1| < 1, |\rho_2| < 1, \ldots, |\rho_{1+f}| < 1$$

Consequently, the acc. surplus variable (S_n) converges to zero for large values of n.

(ii) At least one of the absolute values of the roots is greater than the unity i.e.

$$\text{there exists } i \in \{1,2,\ldots,1+f\} \text{ such that } |\rho_i| > 1$$

consequently, (S_n) diverges to infinity.

(iii) At least one of the absolute values of the roots equals to the unity (while the other are less than unity). Then the (S_n) converges to a certain finite limit determined by the coefficient of the specific root.

Secondly, we must stress again that if the roots are complex numbers the solution exhibits oscillations. So, combining the last statement with the three above options we conclude that there are actual six patterns for the behavior of the solution i.e.

(i) Converging to zero with non-oscillations.

(ii) Converging to finite value with non-oscillation.

(iii) Diverging to infinity with non-oscillation.

(iv) Converging to zero with oscillation.

(v) Converging to finite value with oscillation

(vi) Diverging to infinity with oscillation.

Of course the magnitude of the patterns above are affected by the magnitude and position of the roots in the z-plane. Now, let us consider the characteristic equation (4.11.14) with a slightly different notation.

$$\varphi_1(z) = z^{1+f} - Rz^f + eR\varepsilon = 0$$

and design the paths of the roots using the ten (10) steps of Appendic III (Root Locus Method).

We shall examine the root-locus with respect to parameter (ε). The ten steps below are exactly parallel to Appendix III. Hence,

Step 1: Equation (4.11.14) is rearranged as

$$\varphi_1(z) = 1 + \varepsilon\varphi_2(z) = 0 \tag{4.16.1}$$

$$\varphi_2(z) = \frac{eR}{z^{1+f} - Rz^f} \qquad (4.16.2.)$$

Step 2: The number of separate root loci equals to 1+f.

Step 3: Determine zeros & poles of $\varphi_2(z)$ i.e.

zeros: ∞ (infinity, with 1+f multiplicity)

poles: 0 (with f multiplicity) and R.

Step 4: Complex portion of root-locus.

- f=0 then there is no complex portion.
- f≥1 then the complex portions exist and are symmetrical to x-axis.

Step 5: Real portion of root locus:

- f odd number (1,3,5,7,...) then 1+f is even and consequently no real portion is at left-hand side of zero. The real portion is restricted in the area between 0 and R.
- f zero or even number (0,2,4,6,...) then 1+f is odd and consequently there are real portion at the left hand side of zero and between 0 and R.

Step 6: The angles of the asymptotes a_n

$$a_n = \pm\frac{n\pi}{1+f}, \quad n=1,3,5,7,... \tag{4.16.3}$$

Step 7: The intersection point of asymptotes with x-axis.

$$(IS) = \frac{R}{1+f} \tag{4.16.4}$$

Step 8: Breakpoints (of our locus) with x-axis.

$$\frac{d}{dz}\left[\frac{1}{eR}\left(-z^{1+f} + Rz^{f}\right)\right] = 0 \Leftrightarrow$$

$$z = 0\ \left((f-1)\ \text{multiplicity}\right) \ \&\ z = R\frac{f}{1+f} \tag{4.16.5}$$

Step 9: Not applicable as there are no complex poles.

Step 10: The intersection points may be found after sketching the loci. (So the initial guess is determined by the diagram).

Keeping in mind the distinctions described in steps 4 and 5 we may draw four diagrams for the root-locus i.e.

Diagram (4.16.1): f equals zero (0)

Diagram (4.16.2): f equals one (1)

Diagram (4.16.3): f is an odd number (large value)

Diagram (4.16.4): f is an even number (large value)

Observing the general diagrams (4.16.3) and (4.16.4) we may state the following general comments with respect to the position of the roots and the required stability.

Comments

1) The first important result is revealed by the 3rd step which determines the zeros & poles. One of the poles is R, normally greater than zero. As we observe a path starts at R (which corresponds to ε=0) and goes towards the breakaway point (via the real axis) which may lie left or right to unity.

So, for all small values of ε there is always a root with absolute value greater than unity, consequently the system is unstable. If the path (starting from R) crosses the unity then the specific value of ε is determined by:

$$\varepsilon = \left|\frac{1^{1+f} - R1^{f}}{eR}\right| \tag{4.16.6}$$

(see magnitude criteria of App. III) or equivalently,

$$\varepsilon = \frac{R-1}{eR} \tag{4.16.7}$$

Hence, if $0 < \varepsilon < \dfrac{R-1}{eR}$ (4.16.8)

then the system is unstable (for any value of f).

2) The second important result is revealed by the combination of 3rd and 8th step. Considering the situation before, but assuming that the breakaway point is greater than unity i.e. (from equation 4.16.5).

$$R\frac{f}{1+f} > 1 \Leftrightarrow f > \frac{1}{R-1} \tag{4.16.9}$$

then the certain path lies always outside the unit circle. See Diagram (4.16.4). So if the (4.16.9) inequality holds i.e. the delay factor is greater than a certain value, say f_∞ then the system is unstable independently of the choice of ε.

The critical value of f_∞ may have a very interesting verbal interpretation if we consider R=1+j where j is the interest rate. Then f_∞ equals the perpetuity at interest j

$$f_\infty = a_{\overline{\infty|}}^{(j\%)} \tag{4.16.10}$$

3) Finally if the breakaway point is at left hand side of the unity, it will be interesting to find the intersection points of the unit circle and the root loci which starts from R goes towards the breakaway point and then is developed in two symmetrical (with respect to x-axis) lines which converge asymptotically to two straight lines with

intersection point: $(IS) = \dfrac{R}{1+f}$

angles with x-axis: $(ang)_1 = \dfrac{\pi}{1+f}$, $(ang)_{-1} = -\dfrac{\pi}{1+f}$

Then we calculate the value of ε (say ω) which corresponds to the points above by the magnitude criteria of App. III and conclude: The system will be stable if and only if we choose ε such that

$$\frac{R-1}{\varepsilon \cdot R} < \varepsilon < \omega \qquad (4.16.11)$$

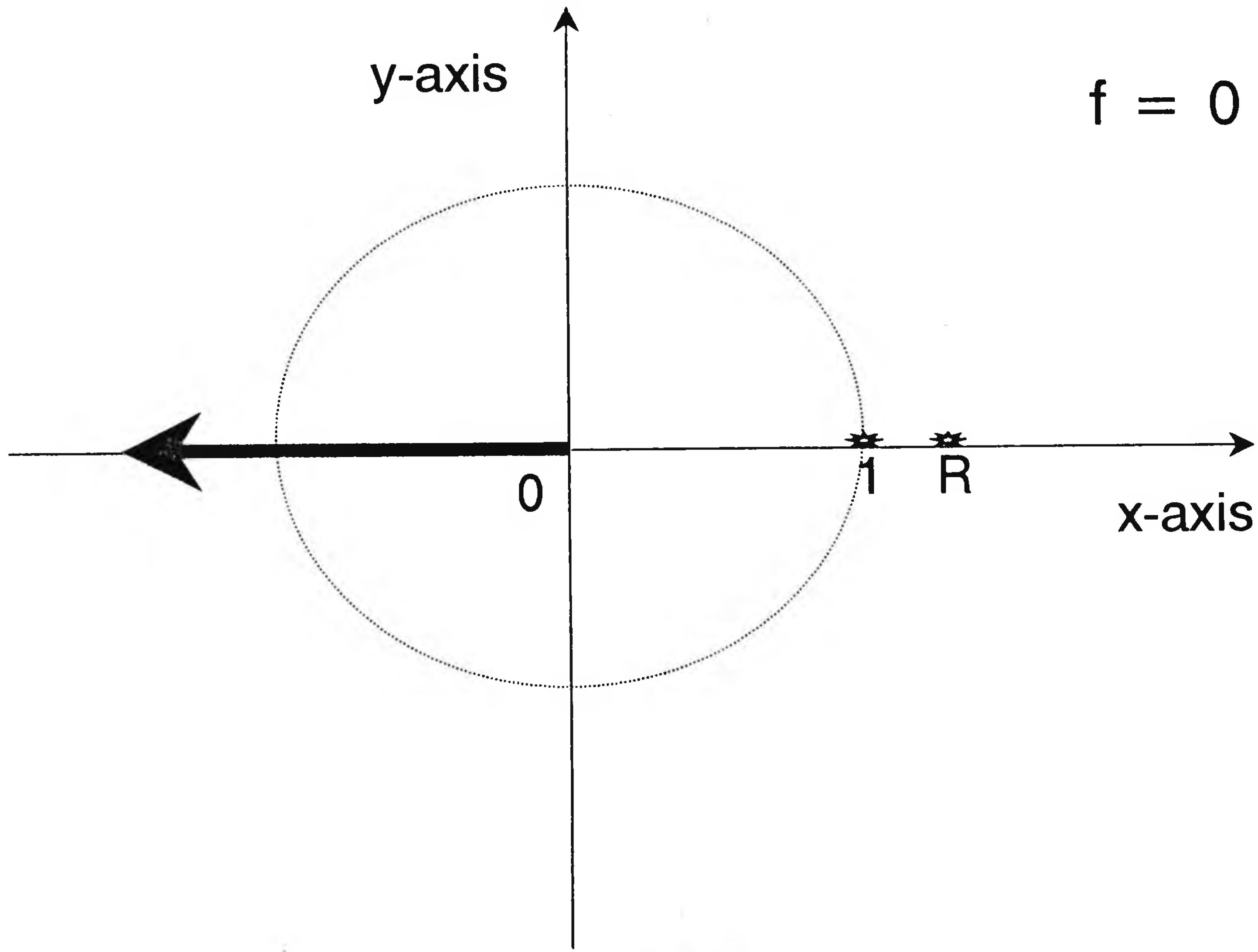


Diagram (4.16.1)

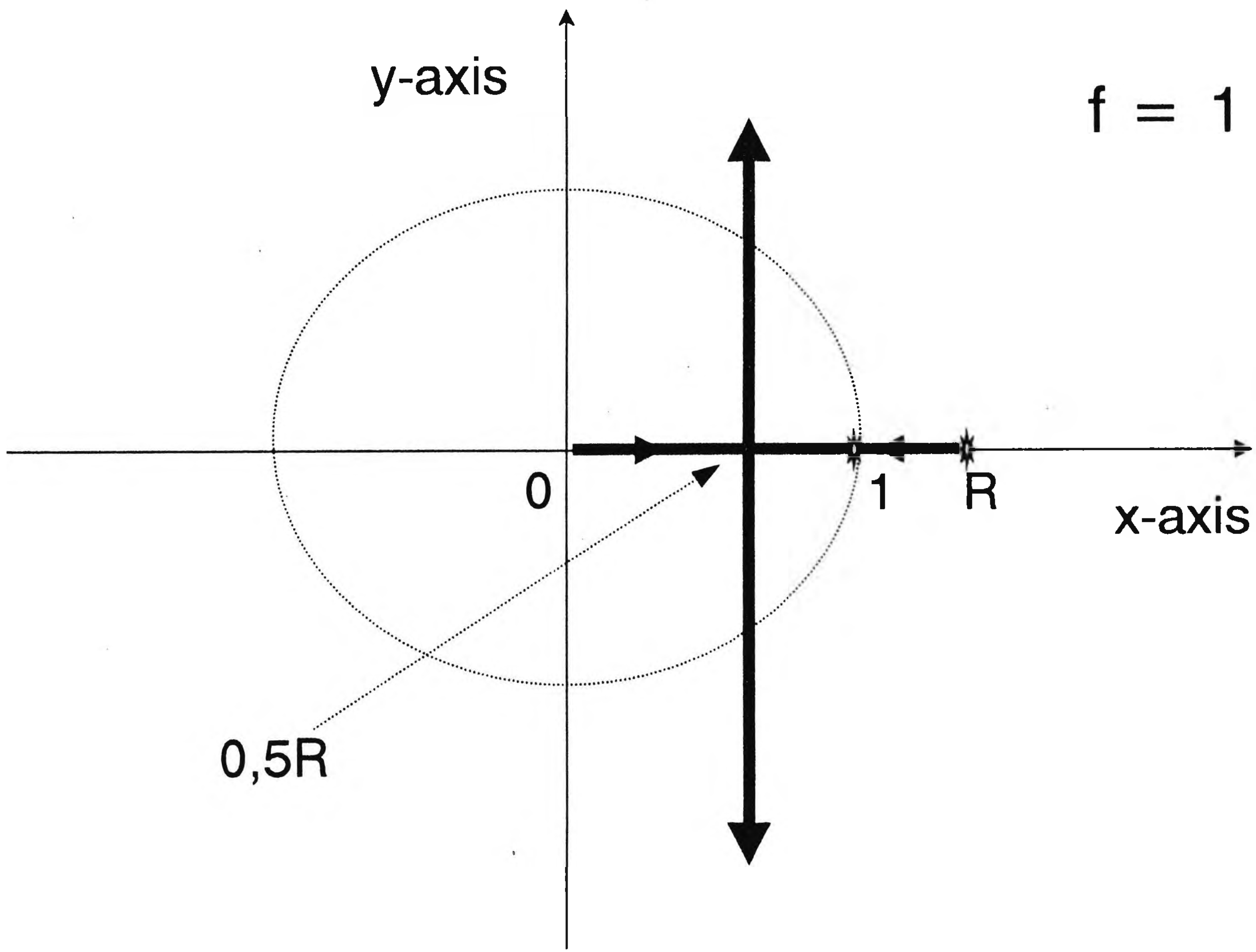


Diagram (4.16.2)

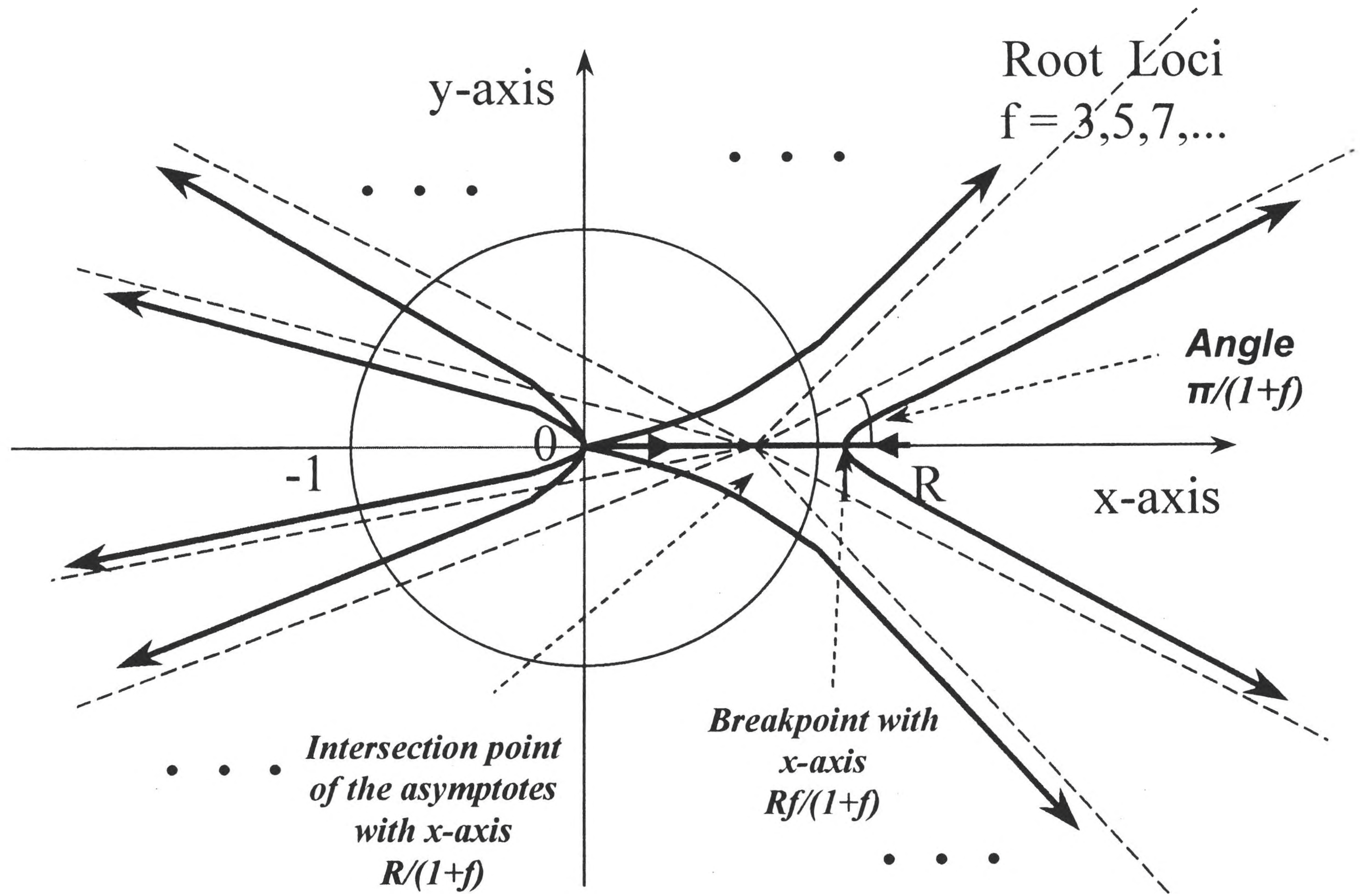


Diagram (4.16.3)

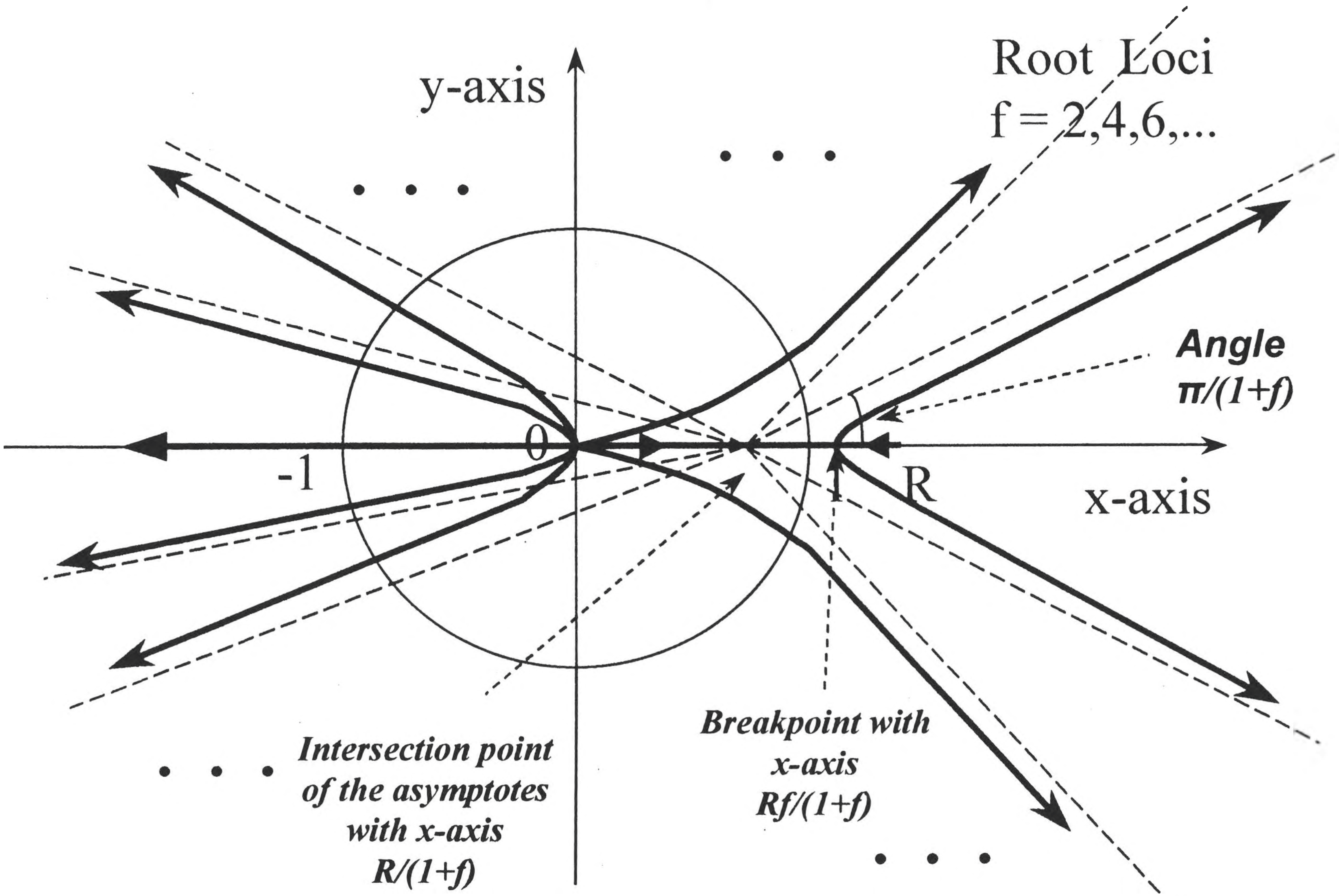


Diagram (4.16.4)

4.17 Stability analysis and optimal design for the parameters of the special case (p=2, f=1, θ=1)

For the special case where f=1 the characteristic equation (4.11.14) becomes

$$z^2 - Rz + e\varepsilon R = 0 \tag{4.17.1}$$

As we can see from diagram (4.16.2) the roots of the equation lies on a cross with the center on the point with coordinates $\left(0, \frac{R}{2}\right)$.

The analytical form of the roots is given below

$$\rho_{1,2} = \frac{R \pm \sqrt{R^2 - 4e\varepsilon R}}{2} \tag{4.17.2}$$

In order to determine the interval of values for ε which results a stable system, we shall use firstly comment (1) of section (4.16) and equation (4.17.2). From the 1st comment we immediately derive that

$$\varepsilon > \frac{R-1}{\varepsilon R} \tag{4.17.3}$$

From equation (4.17.2) and diagram (4.16.2) we shall consider the crossing points with the unit circle (the absolute value will be equal to unity) i.e.

$$|\rho_{1,2}| = 1 \tag{4.17.4}$$

Consequently,

$$\left[\left(\frac{R}{2}\right)^2 + \left(\frac{\sqrt{4e\varepsilon R - R^2}}{2}\right)^2\right]^{\frac{1}{2}} = 1 \Leftrightarrow$$

$$\varepsilon = \frac{1}{eR} \tag{4.17.5}$$

Hence, the system is stable if and only if

$$\frac{R-1}{eR} < \varepsilon < \frac{1}{eR} \tag{4.17.6}$$

As regards the optimal design for the parameters with reference to the fastest return into initial conditions we should try to minimize the magnitude of both roots ρ_1 and ρ_2 as the magnitude of surplus variable depends on them.

The above requirement may be achieved if we have a double root i.e.

$$R^2 - 4e\varepsilon R = 0 \Leftrightarrow \varepsilon = \frac{R}{4e} \tag{4.17.7}$$

So if we select the above specific value the system will converge (in the fastest way) back to initial state (zero).

4.18 Optimal design for the parameters of the special case (p=2, f=2, θ=1)

We shall examine the optimal parameter design with respect to the fastest response of the system for the special case when f=2. The characteristic equation is the following:

$$z^3 - Rz^2 + e\varepsilon R = 0 \tag{4.18.1}$$

If we observe diagram (4.16.4) we can see that

(1) For $\varepsilon=0$ there are two roots **O** (double root) and R.

(2) There are three paths which start (the two of them) from zero (going across the x-axis left and right) and the other from R (going left on the x-axis.

If we demand to minimize the maximum absolute value of the roots then we should choose ε such that the equation has a double root at the breakaway point i.e. $\frac{2R}{3}$ and a third one say ρ on x-axis. Then the following equations hold

$$z^3 - Rz^2 + e\varepsilon R = \left(z - \frac{2R}{3}\right)^2 (z-\rho)$$

$$z^3 - Rz^2 + e\varepsilon R = z^3 - \left(\frac{4R}{3} + \rho\right) z^2 + \left(\frac{4R^2}{9} + \frac{4\rho R}{3}\right) z - \frac{4}{9}R^2\rho$$

Equating the coefficients we obtain

$$\left.\begin{aligned} &\frac{4R}{3} + \rho = R \\ &\frac{4R^2}{9} + \frac{4R}{3}\rho = 0 \\ &-\frac{4}{9}R^2\rho = e\varepsilon R \end{aligned}\right\} \Rightarrow \qquad \begin{aligned} &\rho = -\frac{R}{3} && (4.18.2) \\ &\varepsilon = \frac{4R^2}{27e} && (4.18.3) \end{aligned}$$

So, if we select ε as determined in equation (4.18.3) we shall have the fastest response of the system back to initial conditions.

4.19 The ultimate state of the system – Integral Action

Finally, in this section we shall examine the ultimate state of the system. We shall find the limit of the surplus variable S_n as n (time) goes to infinity.

There is a basic equation which connects the limit of a variable with the limit of the z-transformed respective variable i.e. (see Appendix I).

$$\lim_{n\to\infty} y_n = \lim_{z\to 1} y_z \cdot (z-1)$$

We shall consider the basic equation (4.11.11) which relates the variables C_z and S_z i.e.

$$S_z = -\frac{1 - \frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^p z^{1+f}\right]}{1 - Rz + eR\varepsilon z^{1+f}} C_z$$

and substitute successively the C_z function for a spike and a step input signal.

Spike

As we have seen, the spike function has the form

$$C_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \quad \text{and z-transformation } C_z=1$$

Hence, we obtain

$$\ell_1 = \lim_{n\to\infty} S_n = \lim_{z\to 1} S_z \cdot (z-1) = \lim_{z\to 1} G(z) \cdot C_z (z-1)$$

$$= \lim_{z\to 1} \frac{\frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^p z^{1+f}\right] - 1}{1 - Rz + eR\varepsilon z^{1+f}}(z-1) =$$

$$= \lim_{z\to 1}(z-1) \cdot \lim_{z\to 1} \frac{\frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^p z^{1+f}\right] - 1}{1 - Rz + eR\varepsilon z^{1+f}} = 0 \qquad (4.19.1)$$

i.e. the process will converge to zero when there is a spike as an input signal. This behaviour may be identified also in diagram (4.14.2).

Step

As we have seen the step function has the form, $C_n=1$, $\forall n \in N$ and consequently z-transformation $C_z = \frac{z}{z-1}$.

Hence, we obtain (similarly as above)

$$\ell_2 = \lim_{z\to 1} \frac{\frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^p z^{1+f}\right] - 1}{1 - Rz + eR\varepsilon z^{1+f}} \frac{z}{z-1}(z-1) =$$

$$= \frac{\frac{R}{M}\left[F^{p+f}\theta + \ldots + F^{1+f}\theta^p\right] - 1}{1 - R + eR\varepsilon} \qquad (4.19.2)$$

The last quantity (the value of limit ℓ_2) may be identified also in diagram (4.14.3) as the final distance between each simulation path and x-axis.

As we can observe $\ell_2 \neq 0$, so there is a steady error in the system which may be corrected by using integral action (suggested also by Balzer (1982), Loades (1998)). Integral action is based on integrals (continuous time models) or on summations (discrete time models) of the difference between the actual and the target value of the controlled variables. In our model we shall consider the summation over consecutive years (from 1 up to n-f-1) of the output variable (accumulated surplus S) as the target value for S equals zero. So under proportional & integral action equation (4.11.3) will take the form

$$P_n = \hat{C}_n^2 - \varepsilon S_{n-f-1} - \varepsilon_{ig} \sum_{k=-\infty}^{n-f-1} S_k \tag{4.19.3}$$

and then equation (4.11.12) which describes the transfer function of the system becomes

$$G(z) = -\frac{1 - \frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^{p} z^{1+f}\right]}{1 - Rz + e\varepsilon R z^{1+f} + e\varepsilon_{ig} R \sum_{k=1}^{\infty} z^{f+k}} \tag{4.19.4}$$

As we observe in the denominator exists an infinite summation which equals to (since $|z| < 1$),

$$\sum_{k=1}^{\infty} z^{f+k} = \frac{z^{1+f}}{1-z} \tag{4.19.5}$$

Now if we calculate the ultimate state of the system (under integral action) using the transfer function of (4.19.4) (and with a step input signal) we obtain

$$\ell_3 = \lim_{n\to\infty} S_n = \lim_{z\to 1} G(z)\frac{z}{z-1}(z-1) = \lim_{z\to 1} zG(z) = 0$$

The last limit equals zero because if we combine equations (4.19.4) and (4.19.5) then G(z) may be written in the form

$$G(z) = (z-1)H(z) \qquad \text{where} \qquad H(z) = -\frac{1 - \frac{R}{M}\left[F^{p+f}\theta z^{p+f} + \ldots + F^{1+f}\theta^{p} z^{1+f}\right]}{(z-1)\left[1 - Rz + e\varepsilon R z^{1+f}\right] + e\varepsilon_{ig} R z^{1+f}}$$

So $\qquad \ell_3 = \lim_{z\to 1} G(z) = \lim_{z\to 1} z(z-1)H(z) = 0$

We must stress that all the calculations above exists given that there is a finite limit for y_n. In diagram (4.14.1) we observe that the process diverges although we have an input spike signal. That happens because the combination of the parameter values e,R,ε produce roots ρ_1,ρ_2 with absolute value greater than the unity.

4.20 Analysis of the special case (p=2, f=1, θ=1) using a time varying format for (R_n) and premium delays

In this section, we shall consider the special case where p=2, f=1 and θ=1 (see section (4.13)) with two modifications. The first one will relax the restriction of the interest factor (R_n) from a constant value (R) to the most general where (R_n) may have any pattern of values whether deterministic or even stochastic. The second modification affects the premium. We assume that there is a certain delay in the premium collection procedure.

That may happen in practice as all the policies provide a "grace period" of one month (or even more) to the policy holders in order to pay the required premium. So premiums can not earn the same investment rate as the reserve. Assuming delay ξ' for the premium collection period where $0<\xi'<1$, also define $\xi=1-\xi'$ and use the simple interest model for the accumulation of premiums for the time periods which are less than the unity we can proceed with the new version of our problem.

Under the conditions described above equations (4.11.8) and (4.11.6) will be written respectively as (assuming also that $M' = F^2 + F^3$):

$$S_n = R_n S_{n-1} + \frac{\xi \cdot R_n}{M'}\left[F^3 \cdot C_{n-3} + F^2 \cdot C_{n-2}\right] - e\varepsilon\xi R_n \cdot S_{n-2} - C_n \tag{4.20.1}$$

$$P_n = \frac{1}{M'e}\left[F^3 C_{n-3} + F^2 C_{n-2}\right] - \varepsilon S_{n-2} \tag{4.20.2}$$

The system of equations (4.20.1) & (4.20.2) has a linear time-varying format and may be transformed in the standard vector format of a dynamic system (see section (2.6))

$$\underline{x}_n = \begin{bmatrix} S_n \\ S_{n-1} \end{bmatrix} \in \mathbb{R}^2, \quad \underline{u}_n = \begin{bmatrix} C_n \\ C_{n-1} \\ C_{n-2} \\ C_{n-3} \end{bmatrix} \in \mathbb{R}^4, \quad \underline{y}_n = P_n \in \mathbb{R}$$

$$A_n \in \mathbb{R}^{2\times 2}, \quad B_n \in \mathbb{R}^{2\times 4}, \quad C \in \mathbb{R}^{1\times 2}, \quad D \in \mathbb{R}^{1\times 4}$$

(where $\mathbb{R}$ is the field of the real numbers)

$$A_n = \begin{bmatrix} R_n & -e\varepsilon\xi R_n \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 & \frac{\xi R_n F^2}{M'} & \frac{\xi R_n F^3}{M'} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & -\varepsilon \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 & \frac{F^2}{M'e} & \frac{F^3}{M'e} \end{bmatrix}$$

Hence the dynamic system is written

$$\underline{x}_n = A_n \cdot \underline{x}_{n-1} + B_n \cdot \underline{u}_n \tag{4.20.3}$$

$$\underline{y}_n = C \cdot \underline{x}_{n-1} + D \cdot \underline{u}_n \tag{4.20.4}$$

Now, we shall develop shortly the solution of the time-varying dynamic system.

• n=1, $\underline{x}_1 = A_1 \cdot \underline{x}_0 + B_1 \cdot \underline{u}_1$

• n=2, $\underline{x}_2 = A_2 \cdot \underline{x}_1 + B_2 \cdot \underline{u}_2 =$

$$= A_2\left[A_1\underline{x}_0 + B_1\underline{u}_1\right] + B_2\underline{u}_2 =$$

$$= A_2A_1\underline{x}_0 + A_2B_1 \cdot \underline{u}_1 + B_2\underline{u}_2$$

• n=3, $\underline{x}_3 = A_3\underline{x}_2 + B_3\underline{u}_3 = \ldots =$

$$= A_3A_2A_1\underline{x}_0 + A_3A_2B_1\underline{u}_1 + A_3B_2\underline{u}_2 + B_3\underline{u}_3$$

...

It is easily proved by mathematical induction that

$$\underline{x}_n = \mathscr{A}_n \cdot \underline{x}_0 + \sum_{k=1}^{n} \mathscr{A}_{k+1,n} B_k u_k \quad n=2,3,4,\ldots \tag{4.20.5}$$

where $$\mathscr{A}_{k,n} = \begin{cases} A_k \cdot A_{k+1} \cdot \ldots \cdot A_n & , \quad k < n \\ I & , \quad k > n \end{cases}$$

(Especially when k=1 then $\mathscr{A}_{1,n} \equiv \mathscr{A}_n$).

The $\mathscr{A}_n$ is called the transition matrix and is very important for the behavior of the dynamic system. The solution of the second equation (4.20.4) is easily obtained by substitution of (4.20.5) into it.

$$\underline{y}_1 = Cx_0 + D\underline{u}_1 \tag{4.20.6}$$

$$\underline{y}_n = C\mathscr{A}_{n-1}\underline{x}_0 + \sum_{k=1}^{n-1} \mathscr{A}_{k+1,n-1} B_k u_k + D\underline{u}_n \quad n=2,3,4,\ldots \tag{4.20.7}$$

Closing the section we shall try to analyze the transition matrix $\mathscr{A}_n$ and make some comments about its behaviour.

The calculation of the $\mathcal{A}_n$ matrix is not standard.

Of course we have always the ability of the calculation of $\mathcal{A}_n$ using a computer but in that case we do not have the deep insight into the problem which may be available by the analytical formula.

The analytical calculation of $\mathcal{A}_n$ requires the full knowledge of the sequence of the accumulation factors $\{R_n: n=1,2,3,...\}$. A possible pattern for R_n may follow the "investment cycle" i.e. A sine wave pattern with a certain period where the R_n takes discrete constant values on the sine wave graph e.g. period of the sine wave = 8 years.

n	1	2	3	4	5	6	7	8	9	10
R_n	4%	4.25%	4.5%	4.25%	4%	3.75%	3.5%	3.75%	4%	4.25%

Under the pattern described above, $\mathcal{A}_n$ depends on the product $\mathcal{A}_8=A_1 \cdot A_2 \cdot ... \cdot A_8$. Consequently the investigation of the behavior of the solution of the dynamic system requires the determination of the eigenvalues of $\mathcal{A}_8$.

Another possible pattern for $\{R_n: n=1,2,...\}$ is the following

$$R_n = \begin{cases} R_1 & n = 2k+1 \quad k=0,1,2,... \\ R_2 & n = 2(k+1) \quad k=0,1,2,... \end{cases}$$

i.e. R_n fluctuates between R_1 and R_2 under a strict format (R_1 for odd numbers and R_2 for even numbers).

Then we obtain

$$\mathcal{A}_n = \begin{cases} \mathcal{A}_1 \cdot \mathcal{A}_2^{\frac{n-1}{2}} & \text{if } n = 2k+1 \quad k=0,1,2,... \\ \mathcal{A}_2^{\frac{n}{2}} & \text{if } n = 2(k+1) \quad k=0,1,2,... \end{cases}$$

where

$$\mathcal{A}_2 = \begin{bmatrix} R_2 & -e\varepsilon \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} R_1 & -e\varepsilon \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} R_1 R_2 - e\varepsilon & -e\varepsilon R_2 \\ R_1 & -e\varepsilon \end{bmatrix}$$

Now, in order to obtain the analytical solution of the system we should obtain the eigenvalues of $\mathcal{A}_2$. Actually, in the real world the patterns of R_n may include a stochastic component which makes the analytical approach of the solution more difficult.

In the next section, we shall discuss the concept of stability and produce general result even for the cases where R_n is stochastic.

4.21 Stability analysis of the time-varying format of the model

In this section we shall try to answer the question of the stability for the time-varying format of the system and consequently answer the same question for the special case described in section (4.20).

First of all, we shall find the equilibrium points $\underline{x}_e$. According to definition (2.15.1)

$\underline{x}_e$ equilibrium point $\Leftrightarrow$ $\underline{x}_n = \underline{x}_e$ $\forall n > 0$ given $\underline{x}_0 = x_e$ and $\underline{u}_n = \underline{0}$ $\forall n > 0$

or equivalently that $\underline{x}_e$ is the solution of the homogeneous system (4.21.1) with initial condition $\underline{x}_0 = \underline{x}_e$

$$\underline{x}_e = \mathcal{A}_n \cdot \underline{x}_e \qquad \forall n \geq 0 \tag{4.21.1}$$

This equation has always the trivial solution $\underline{x}_e = \underline{0}$. Hence, the time-varying system has at least the zero vector as equilibrium point. We shall not investigate for any other equilibrium point as we shall prove that zero is an asymptotic stable point and consequently is the only stable point.

Before examining the asymptotic stability we shall define the norm functions for the vector and matrix spaces as follows:

Let $A \in R^{n\times n}$ then $|||A|||=\rho(A)$ where is the maximum absolute value of the eigenvalues of matrix A. We also define a compatible norm for the vector space $\mathbb{R}^n$ say $||\underline{x}||$. (Compatible means, $||A\cdot x|| \le |||A|||\cdot||x||$).

Now, we may derive the eigenvalues of the A_n matrices as follows

$$|\rho_n I - A_n| = 0 \Leftrightarrow \begin{vmatrix} \rho_n - R_n & e\varepsilon\xi R_n \\ -1 & \rho_n \end{vmatrix} = 0 \Leftrightarrow$$

$$\rho_n^2 - R_n\rho_n + e\varepsilon\xi R_n = 0 \Leftrightarrow$$

$$\rho_n = \frac{R_n \pm \sqrt{R_n^2 - 4e\varepsilon\xi R_n}}{2} \tag{4.21.2}$$

Hence
$$|||A_n||| = \rho(A_n) = \begin{cases} \dfrac{R_n + \sqrt{R_n^2 - 4e\varepsilon\xi R_n}}{2}, & e < \dfrac{R_n}{4e\xi} \\ \sqrt{e\varepsilon\xi R_n} & \varepsilon > \dfrac{R_n}{4e\xi} \end{cases} \tag{4.21.3}$$

(if $\varepsilon > \frac{R_n}{4e\xi}$ then we have complex roots and $|\rho_n| = \sqrt{e\varepsilon}$ for any ρ_n).

Now, considering the definition (2.15.2) about the Liapunov stable point we have the equivalent condition. $\underline{x}_e$ is Liapunov stable point $\Leftrightarrow$ $\exists K > 0 : |||\mathscr{A}_n||| \le K \quad \forall n > 0$ (i.e. the transition matrix is bounded for any n>0).

$$|||\mathscr{A}_n||| = |||A_1 \cdot A_2 \cdot \ldots \cdot A_n||| \le |||A_1|||\cdot|||A_2|||\cdot\ldots\cdot|||A_n||| =$$

$$= \rho(A_1)\cdot\rho(A_2)\cdot\ldots\cdot\rho(A_n) \le [\rho_{n,max}]^n$$

where $\rho_{n,max} = \max\{\rho(A_k)\ ,\ k = 1,2,\ldots,n\}$ and $\rho_{max} = \max\{\rho(A_k)\ ,\ k = 1,2,\ldots\}$

and as we can see from equation (4.21.2) the ρ_{max} is obtained for the maximum R_n. Hence if we consider $R_{max} = \max\{R_n = 1,2,...,\}$ (if any) we obtain

$$\rho_{max} = \begin{cases} \dfrac{R_{max} + \sqrt{R_{max}^2 - 4e\varepsilon\xi R_{max}}}{2}, & \varepsilon < \dfrac{R_{max}}{4e\xi} \\ \sqrt{e\varepsilon,\xi R_{max}} & \varepsilon > \dfrac{R_{max}}{4e\xi} \end{cases} \tag{4.21.4}$$

Now if we choose ε such that $\rho_{max}<1$ then

$$\left|\left|\left|\mathscr{A}_n\right|\right|\right| \le \left[\rho_{n,max}\right]^n \le \left(\rho_{max}\right)^n \le 1 \qquad \forall n > 0$$

so the transition matrix is bounded and consequently $\underline{x}_e = \underline{0}$ is Liapunov stable. Furthermore we may show that $\underline{0}$ is an asymptotic stable point and consequently the only equilibrium point. We must show that

$$\lim_{n\to\infty}\left\|\underline{x}_n - \underline{0}\right\| = 0$$

Again we choose ε such that $\rho_{max}<1$ then

$$\lim_{n\to\infty}\left\|\underline{x}_n\right\| = \lim_{n\to\infty}\left\|\mathscr{A}_n \cdot \underline{x}_0\right\| \le \lim_{n\to\infty}\left|\left|\left|\mathscr{A}_n\right|\right|\right| \cdot \left|\left|\left|\underline{x}_0\right|\right|\right| \le$$

$$\left\|\underline{x}_0\right\| \cdot \lim_{n\to\infty}\left(\rho_{max}\right)^n = \left\|\underline{x}_0\right\| \cdot 0 = 0$$

Hence $\underline{0}$ is asymptotic stable.

Now, the investigation of ε in order to obtain an asymptotic stable system is restricted to the question of finding an ε such that $\rho_{max}<1$. From equation (4.21.4) we obtain the system of inequalities

$$\frac{R_{max} + \sqrt{R_{max}^2 - 4e\varepsilon\xi R_{max}}}{2} < 1 \tag{4.21.5}$$

$$\sqrt{e\varepsilon\xi R_{max}} < 1 \tag{4.21.6}$$

Solving the system we obtain

$$\frac{R_{max}-1}{e\xi R_{max}} < \varepsilon < \frac{1}{e\xi R_{max}} \tag{4.21.7}$$

Closing this section we shall summarize our results under the following comments.

Comments

Consider Ω as the set where R_n takes its values (under a deterministic or a stochastic model) then:

(1) If Ω is bounded then our system is Liapunov stable.

(2) If Ω is bounded, we may choose ε as

$$\frac{R_b-1}{e\xi R_b} < \varepsilon < \frac{1}{e\xi R_b}$$

(where R_b is the lower upper bound of Ω)

and force the system to be asymptotic stable to zero.

(3) If Ω is not bounded then system is unstable.

Numerical Example

We shall consider now, a numerical example to illustrate in a better way the ideas expressed in this section. For this reason we assume the following:

(1) Premium delay (ξ')	:	0 (No delay, so ξ=1)
(2) Expense factor (e)	:	80%
(3) Inflation factor (F)	:	1.04
(4) Input signal	:	Step (0,1,1,...)

(5) Interest factor (R_n) : Uniform in [108, 1.10] i.e. the interest factor follows a stochastic time varying pattern for all the years n=1,2,3,...

From the last assumption we obtain that

$$R_b=1.10$$

So the bounds for (ε) the feedback factor using inequality (4.21.7) in order to produce a stable system are defined below i.e.

$$0.136 < \varepsilon < 1.136 \tag{4.21.8}$$

In the next table (4.21.1) and the respective diagram (4.21.1) we can observe that the first path corresponding to ε=5% diverges while all the others (making true the expression (4.21.8)) converge to certain values.

Other numerical examples may be also considered using different patterns for the interest factor (R_n). A lognormal distribution would be a natural choice for the (R_n).

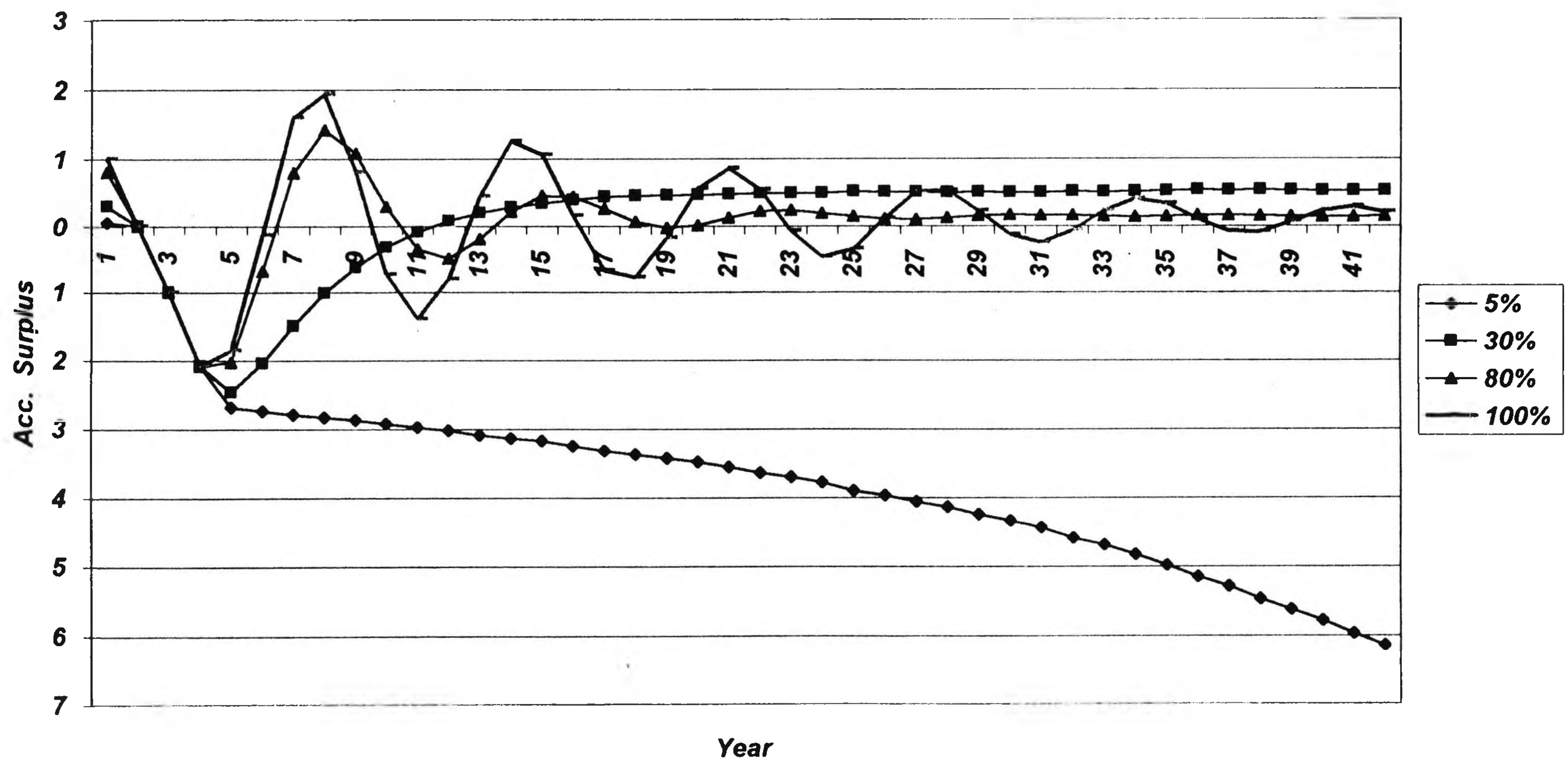
TABLE (4.21.1)

Input Signal *Step* *(0,1,1,1,1,1,...)*

Expense (e) *80%*
Interest (R_n) *Uniform [1,08 1,10]*
Infflation (F) *1,04*
(M/) *2,206*

			(I)	(ii)	(iii)	(iv)
Feedback (ε)			**5%**	**30%**	**80%**	**100%**
Time	**Interest Factor (R_n)**	**Claims**	**Accumulated surplus**	**Accumulated surplus**	**Accumulated surplus**	**Accumulated surplus**
0	0	0	0,0000	0,0000	0,0000	0,0000
1	1,0993	1	-1,0000	-1,0000	-1,0000	-1,0000
2	1,0843	1	-2,0843	-2,0843	-2,0843	-2,0843
3	1,0867	1	-2,6888	-2,4714	-2,0368	-1,8629
4	1,0840	1	-2,7403	-2,0528	-0,6779	-0,1279
5	1,0984	1	-2,7934	-1,5049	0,7856	1,5948
6	1,0877	1	-2,8314	-1,0133	1,4140	1,9336
7	1,0865	1	-2,8684	-0,6220	1,0766	0,8012
8	1,0970	1	-2,9254	-0,3186	0,2852	-0,7211
9	1,0936	1	-2,9800	-0,0916	-0,3480	-1,3959
10	1,0884	1	-3,0278	0,0720	-0,4890	-0,8030
11	1,0969	1	-3,0935	0,1999	-0,1952	0,4410
12	1,0838	1	-3,1376	0,2818	0,2114	1,2579
13	1,0816	1	-3,1782	0,3344	0,4454	1,0606
14	1,0970	1	-3,2518	0,3897	0,4371	0,1565
15	1,0937	1	-3,3238	0,4321	0,2601	-0,6631
16	1,0825	1	-3,3748	0,4491	0,0612	-0,7709
17	1,0837	1	-3,4296	0,4580	-0,0303	-0,1767
18	1,0818	1	-3,4822	0,4607	0,0066	0,5577
19	1,0916	1	-3,5599	0,4745	0,1200	0,8548
20	1,0926	1	-3,6448	0,4903	0,2191	0,5391
21	1,0804	1	-3,7037	0,4871	0,2342	-0,0759
22	1,0877	1	-3,7822	0,4895	0,1899	-0,4640
23	1,1000	1	-3,8973	0,5098	0,1439	-0,3436
24	1,0801	1	-3,9659	0,5038	0,1043	0,1099
25	1,0886	1	-4,0590	0,5039	0,1018	0,5075
26	1,0810	1	-4,1352	0,4949	0,1189	0,5345
27	1,0907	1	-4,2424	0,4986	0,1493	0,2308
28	1,0839	1	-4,3351	0,4955	0,1632	-0,1294
29	1,0857	1	-4,4365	0,4937	0,1591	-0,2553
30	1,0979	1	-4,5825	0,5094	0,1579	-0,0688
31	1,0825	1	-4,6859	0,5056	0,1432	0,2291
32	1,0919	1	-4,8245	0,5105	0,1379	0,4021
33	1,0940	1	-4,9788	0,5197	0,1446	0,3334
34	1,0965	1	-5,1513	0,5320	0,1583	0,1093
35	1,0855	1	-5,2902	0,5277	0,1569	-0,0853
36	1,0941	1	-5,4686	0,5318	0,1550	-0,0949
37	1,0853	1	-5,6200	0,5250	0,1445	0,0564
38	1,0860	1	-5,7798	0,5175	0,1352	0,2297
39	1,0907	1	-5,9684	0,5178	0,1374	0,2921
40	1,0863	1	-6,1458	0,5138	0,1415	0,2039

Diagram (4.21.1)
Stochastic pattern for the (R_n) Interest Factor



4.22 A measuring element for the entropy of an insurance system

The concept described by the Greek word «Entropy» (or in the Greek version «Entropia») is widely used for all physical systems. The measurement of this quantity is quite important for every system as it determines the order («taxis») or the respective chaos in it. It also provides the potential ability of the system to return back to its initial state.

Clearly from the last definition, as the delay factor increases the entropy of the system increases too, because the root loci which lies within the unit circle are reduced i.e. there are less potential choices for (ε) feedback factor in order to have a stable system. It is also clear that if f is greater than the critical value of f_∞ then the system entropy goes to infinity as there is no chance to return back to the initial state.

Diagrammatically (with respect to the (f) delay factor) we may think the graph of entropy starting from zero (when f=0) and going quickly to infinity (when $f=f_\infty$) and remaining thereafter at this state.

Since entropy is analogous to the area of root loci lying in the unit circle we may have the following measuring elements (for the entropy).

(1) Angle of the first asymptote with x-axis

$$\text{Angle} = \frac{\pi}{1+f}$$

The larger the angle the less entropy of the system. When the angle reduces then the root loci quickly goes out of the unit circle and entropy increases.

(b) The breakaway point of the root loci with x-axis is equal to

$$R\frac{f}{1+f}$$

Obviously the greater R the greater the value for the breakaway point and consequently the less this root loci-within the unit circle. In this case we obtain large values for the entropy of the system.

(c) A more sophisticated measuring element may be described as

$$\text{Entropy} = \frac{1}{\psi}$$

when ψ corresponds to the length of the interval $[\psi_1,\psi_2]$ where $[\psi_1,\psi_2]\subseteq[0,1]$ and for any value of (ε) in the $[\psi_1,\psi_2]$ the respective roots (solution of the system) lie within the unit circle.

The left boundary of the interval $[\psi_1,\psi_2]$ is determined from the equation

$$f(1) = 1^{1+f} - R\cdot 1^{f} + e\varepsilon R = 0$$

Hence $$\varepsilon = \psi_1 = \frac{R-1}{eR}$$

The calculation of ψ_2 requires numerical approximations.

Finally after all this analysis we may say that great concern should be paid when designing a system in order to obtain less entropy.

4.23 Conclusions

In this final section, closing Chapter 4, we shall briefly review the basic modelling features and the important results.

Our problem may well be described with the title of this chapter, i.e. «Insurance Pricing».

As we have stated in the introductory section the basic model is similar with that developed by Balzer & Benjamin (1980). Actually it should be considered as a full generalization with respect to the delay factor (f) which is left as a free control parameter into the model.

The model is also similar to that of Vandebroek & Dhaene (1990) who investigate the optimal premium control while we are focusing on stability and optimal design of the parameters of the system (and especially the feedback factor). The most important result is the existence of the critical value for (f)

$$f_\infty = \frac{1}{R-1}$$

If we select $f > f_\infty$ then the process will diverge to infinity, independently of the choice of feedback factor (ε).

As we have also stated, f_∞ equals to the perpetuity in arrears i.e.

$$f_\infty = a_{\overline{\infty|}}^{(j\%)}$$

Considering some values of j% we obtain respectively the values of f_∞ i.e.

j%	0%	5%	10%	20%	25%	50%	100%
f_∞	∞	20	10	5	4	2	1

For an extra large value of (R) as R=100% we obtain f_∞=1 which means that the system will always diverge (of course except the case where f=0). The divergency of the process independently of the choice of (ε) means that the information included in the surplus variable with delay (f) (i.e. S_{n-f}) is needless to the system. Hence, large values of (R) (or equivalently j%) deteriorate quickly (in f_∞ time units) the information carried by the surplus variable. At the other extreme where j=0% (so R=1) we obtain $f_\infty=\infty$ i.e. we may always find a value of (ε) in order to stabilize the system.

Another interesting result is the optimal design for parameters involved when f=1 or f=2 (sections (4.17), (4.18) i.e. The fastest response of the output (return back to initial position) is obtained:

for f=1 when $\varepsilon = \dfrac{R}{4e}$ and

for f=2 when $\varepsilon = \dfrac{4R^2}{27e}$

The second direction of full generalization is the introduction of the time-varying pattern for interest factor (R) (whether deterministic or stochastic). The basic result is connected with the bound of the Ω set (where Ω is the set of possible values of R_n, n=1,2,...).

Finally, the third generalization is the incorporation of a more advanced formula as a claim predictor using an inflation factor F, a weighting factor θ and averaging over (p) years. (Where (p) is another control parameter.) At this point, we recall a result from simulations in section (4.14) for the random input signal where a value ε=50% or 60% appears to be the optimal ones (with respect to the smaller variance of the output).

Chapter 5

Application to Multinational Pooling Arrangements

5.1 Introduction

In the real world, there are many types of insurance agreements operating amongst two, three or more insurance companies. Actually these are types of reinsurance agreements, which determine the way of interaction amongst the surplus funds of the companies.

In this chapter we shall focus on a special reinsurance agreement known as "multinational pooling". The concept of pooling is one of the efficient tools for handling large employee benefit schemes of international companies.

Furthermore, we shall design a model of pooling on a special quota share basis with respect to the surplus fund of each company using control techniques and we shall investigate the potential dynamic behavior of the system.

We shall present the problem in the most general form (i.e. assuming (m) companies and the existence of a delay factor (f)) but provide analysis for three special cases i.e.

1st: m not fixed with f=1

2nd: m=2 with f not fixed

3rd: m=2 with f=1

The proposed problem also deals partially with the question of the optimal reinsurance as apppears in the relevant literature (section (5.19)). The literature tackles the problem from a static point of view while our approach concentrates on the dynamics of the system.

Before we go on with the typical formulation of the equations we shall briefly discuss the concept of pooling and the various multinational insurance networks.

5.2 Description of Multinational Insurance Networks

The international industrial giants of the last decades or the other large and medium sized corporations operating in many countries all over the world have created a special insurance market with some specific needs.

Operating on a centralized or decentralized basis the parent corporations have demanded some kind of control on their subsidiaries. The insurance companies have quickly received the message and gave their answer through the construction of Multinational Insurance Networks.

These networks have initially been established through special reinsurance agreements between affiliated insurance companies. Nowdays, at the end of the century, the traditional networks have been collapsed and gave their position to the modern networks which are fully controlled (and owned) by the insurance giants of the world market.

The basic concept and perhaps the most important product which is being sold through these networks is the "pooling arrangement". Generally speaking, pooling is a special kind of self insurance. More precisely we may say that: Pooling is the combination of risks underlying the employee benefit schemes in two or more countries of a certain corporation.

A formal way to establish a pooling arrangement is the following:

- A large multinational corporation (e.g. IBM computer company) goes to the Central Office of the network asking for insurance coverage for their employees around the world.
- The Central Office informs the insurance companies of the network and the parent multinational corporation informs its local subsidiaries for the potential co-operation.
- Each insurance company (coordinated by the Central Office) establishes a group policy with each subsidiary using the local rates and insurance practice.
- The problem which then arises is the coordination of the network i.e. Establish a general procedure for

  (a) Premium rating by sharing the international claim experience of each company in the pool.

  (b) The potential interaction of the surpluses amongst the insurance companies participating in this agreement.

  (The mathematical formulation of the problem will be developed in the next section).

The basic advantages of the "pooling" concept and the respective motivation are listed below:

- **Cost savings:** The large volume of business will result lower expense margins and perhaps lower security loadings as the random fluctuation of the claim experience will be restricted down to the minimum levels. Lower expenses may be realized because of minor efforts (from insurer's side) with respect to acquisition costs, underwriting and claim management.

- **Extra Cost Saving:** The pooling procedure (i.e. the combination of all risks in different countries) will normally produce some extra profit as the parent corporation can benefit from the good claim experience on a larger scale.
- **Flexible Scheme:** Expatriate problems or other complications arising from the internal structure of the corporation may be solved in the most efficient manner.
- **International Control:** As a last benefit (and perhaps the most important) we should put the potential of international control given to the parent corporation. This may result the right decision-making targeting the ideal risk management.

A full description of the multinational insurance networks plus the description of the additional advantages of multinational pooling is provided in the manual of William M. Mercer International (1988).

5.3 Formulation of the control problem for the general model of (m) companies with (f) as a delay factor

As we have stated in the introductory section we shall examine a special quota share agreement (with respect to the surplus funds) of a pooling arrangement with the following characteristics.

(a) There are (m) insurance companies which participate in the multinational network and

(b) Each company passes to the other block of (m-1) companies a pre-determined percentage of its accumulated surplus at every time n (n=1,2,3,...).

In the most general case, the predetermined percentage which each company passes to the other is not equally divided. So we have to define a matrix Λ described as the "harmonization matrix" which governs the surplus exchange. So,

$$\Lambda = [\lambda_{ij}] \in \mathbb{R}^{m \times m}$$

where λ_{ij} is the predetermined percentage of surplus which the i-th company passes to j-th company.

Obviously, each row adds to unity i.e.

$$\sum_{j=1}^{m} \lambda_{ij} = 1 \qquad \text{for each } i=1,2,\ldots,m$$

and the element of the first main diagonal λ_{ii} i=1,2,...,m determines the retained surplus by each company.

(c) Each company has its own operational parameter values for expense, feedback, accumulation and inflation factors (all of them will be defined properly later in this section).

Definition of the problem

The existing problem may be described as follows: Design, check and establish the best (i.e. stable, controllable, observable and consequently optimal process) strategy for premium rating (control feedback, expenses etc.) and surplus exchange (control Λ matrix) for a combined system of group policies covered under (m) different insurance companies which operate under a special pooling reinsurance agreement.

The development of the model will be based on the model described in Chapter 4 (see the description of the problem section (4.2)). The problem there, refers to a non-life portfolio but it may also adequately describe a situation of a group profit sharing policy. The structure of the current model is a generalization of the model in chapter 4 (which may be treated as a subsystem in the new format of the model).

Now, the pooling arrangement demands a certain combination, multiplicating the parameters of the model. The interesting new problem which now appears is the **interaction** phenomenon between the surpluses of the (m) companies. Consequently, we have to control not only the profit sharing factors denoted by ε_i's but also the interaction denoted with the harmonization matrix Λ. The control action on the Λ matrix will result the desired **harmonization** between the (m) companies.

As a next step we shall define all the required symbols using two subscripts where needed (the first indicating the subsystem and the second indicating the time).

Parameters & Notation

$C_{\ell,n}$: Total amount of annual incurred claims (for the ℓ-th company in the n-th year).

$P_{\ell,n}$: Total gross (including expenses) premium paid at the end (or equivalently there is no investment income from the premiums), of year for the ℓ-th company $\ell=1,2,...,m$ in the n-th year. The assumption that the premium is pai,nd at the end of the year is also reasonable for this model due to the normal network's practise to credit no investment income on premium payments.

e_ℓ : Expense factor $[(1\text{-}e_l)P_{l,n}$ margin for expenses] for the ℓ-th company. So we have a vector of m-dimension for the expense factor, say $\underline{e}=\left[e_1e_2...e_m\right]$.

$L_{\ell,n}$: Annual surplus at the end of the n-th year for the ℓ-th company

$$L_{l,n} = e_l \cdot P_{l,n} - C_{l,n} \tag{5.3.1}$$

R_l : Accumulation factor ($R_\ell = 1 + j_\ell$ rate of return) for the ℓ-th company. The respective vector for accumulation factor is $\underline{R}=\left[R_1R_2...R_m\right]$

λ_{ij} : i,j=1,2,...,m. Constitutes the harmonization matrix Λ (as defined earlier).

$S_{\ell,n}$: Accumulated surplus at the end of the n-th year for the ℓ-th company.

$$S_{\ell,n} = R_1 \cdot \left[\lambda_{1\ell} S_{1,n-1} + \ldots + \lambda_{2\ell} S_{2,n-1} + \ldots + \lambda_{ml} S_{m,n-1}\right] + L_{\ell,n}, \qquad \ell = 1,2,\ldots m \tag{5.3.2}$$

θ_ℓ : Weighting factor of the $\ell-$th company. (As described in chapter 4) $\ell = 1,2,\ldots,m$.

F_ℓ : Inflation factor (F_ℓ=1+inflation rate) of the l-th company. This factor indicates a certain internal growth of the total annual claims which may be attributable to inflation or to business growth. The respective vector for inflation is $\underline{F} = [F_1 F_2 \ldots F_m]$.

$\hat{C}_{\ell,n}$: Estimate of the total expected annual incurred claims in year n (at the beginning of year n i.e. at time n-1). Assuming a delay of information of (f) years and averaging over p consecutive years we obtain

$$\hat{C}_{\ell,n} = \frac{1}{M_\ell} \cdot \left(F_l^{p+f} \theta_\ell C_{l,n-p-f} + \ldots + F_\ell^{1+f} \theta_\ell^p C_{\ell,n-f-1}\right) \qquad \ell = 1,2,\ldots,m \tag{5.3.3}$$

where
$$M_\ell = \sum_{k=1}^{p} F_l^{p+f+1-k} \cdot \theta_\ell^k \tag{5.3.4}$$

ε_ℓ : Profit sharing factor (feedback factor) for the ℓ-th company which includes both the local and international premium repayments and determines the percentage of accumulated surplus repaid. The respective vector is $\underline{\varepsilon} = [\varepsilon_1 \varepsilon_2 \ldots \varepsilon_m]$.

Using all the notation above we may define the control law of our system which determines the premiums for subsequent years i.e.

$$P_\ell = \frac{1}{e_\ell}\hat{C}_{\ell,n} - \varepsilon_\ell S_{\ell,n-f-1}, \quad \ell = 1,2,\ldots,m \tag{5.3.5}$$

According to the description of the model and combining the existing relationships between the parameters (i.e. equations (5.3.1), (5.3.2), (5.3.3), (5.3.4) and (5.3.5)) we obtain a system of (2m) equations.

$$\left.\begin{aligned}
S_{1,n} &= R_1\lambda_{11}S_{1,n-1} + R_1\lambda_{21}S_{2,n-1} + \ldots + R_j\lambda_{m1}S_{m,n-1} + \\
&\frac{F_1^{p+f}\theta_1}{M_1}C_{1,n-f-p} + \ldots + \frac{F_1^{1+f}\theta_1^p}{M_1}C_{1,n-f-1} - e_1\varepsilon_1 S_{1,n-f-1} - C_{1,n} \\
&\vdots \\
S_{m,n} &= R_m\lambda_{1m}S_{1,n-1} + R_m\lambda_{2m}S_{2,n-1} + \ldots + R_m\lambda_{mm}S_{m,n-1} + \\
&\frac{F_m^{p+f}\theta_m}{M_m}C_{m,n-f-p} + \ldots + \frac{F_m^{1+f}\theta_1^p}{M_m}C_{m,n-f-1} - e_m\varepsilon_m S_{m,n-f-1} - C_{m,n} \\
P_{1,n} &= \frac{F_1^{p+f}\theta_1}{M_1 e_1}C_{1,n-f-p} + \ldots + \frac{F_1^{1+f}\theta_1^p}{M_1 e_1}C_{1,n-f-1} - \varepsilon_1 S_{1,n-f-1} \\
&\vdots \\
P_{m,n} &= \frac{F_m^{p+f}\theta_m}{M_m e_m}C_{m,n-f-p} + \ldots + \frac{F_m^{1+f}\theta_m^p}{M_m e_m}C_{m,n-f-1} - \varepsilon_m \cdot S_{m,n-f-1}
\end{aligned}\right\} \tag{5.3.6}$$

The system above may take the standard format of a dynamic system

$$\left.\begin{aligned}
\underline{x}_n &= A\underline{x}_{n-1} + B\underline{u}_n \\
\underline{y}_n &= C\underline{x}_{n-1} + D\underline{u}_n
\end{aligned}\right\} \tag{5.3.7}$$

$$\underline{x}_n = \begin{bmatrix} S_{1,n} \\ S_{1,n-1} \\ \vdots \\ S_{1,n-f} \\ --- \\ S_{2,n} \\ S_{2,n-1} \\ \vdots \\ S_{2,n-f} \\ --- \\ \vdots \\ --- \\ S_{m,n} \\ S_{m,n-1} \\ \vdots \\ S_{m,n-f} \end{bmatrix} \in \mathbb{R}^{m(1+f)}, \underline{y}_n = \begin{bmatrix} P_{1,n} \\ P_{2,n} \\ \vdots \\ P_{m,n} \end{bmatrix} \in \mathbb{R}^m, \underline{u}_n = \begin{bmatrix} C_{1,n} \\ C_{1,n-1} \\ \vdots \\ C_{1,n-f-p} \\ --- \\ C_{2,n} \\ C_{2,n-1} \\ \vdots \\ C_{2,n-f-p} \\ --- \\ \vdots \\ --- \\ C_{m,n} \\ C_{m,n-1} \\ \vdots \\ C_{m,n-f-p} \end{bmatrix} \in \mathbb{R}^{m(1+f+p)}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix} \in \mathbb{R}^{m(1+f)\times m(1+f)}$$

i.e. is a matrix of matrices $A_{ij} \in \mathbb{R}^{(1+f)\times(1+f)}$ where

$$A_{ii} = \begin{bmatrix} R_i\lambda_{ii} & 0 & 0 & \cdots & 0 & -e_i\varepsilon_i \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \underset{(i\neq j)}{A_{i,j}} = \begin{bmatrix} R_i\lambda_{ji} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{bmatrix} \in \mathbb{R}^{m(1+f)\times m(1+f+p)}$$

i.e. B is a matrix of matrices $B_{ij} \in \mathbb{R}^{(1+f)\times(1+f+p)}$ where

$$B_{ii} = \begin{bmatrix} -1 & 0 & \cdots & 0 & \dfrac{F_i^{1+f}\theta_i^p}{M_i} & \cdots & \dfrac{F_i^{p+f}\theta_i}{M_i} \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}, \; B_{ij} = O \qquad i \neq j$$

$$C = [C_1 C_2 \ldots C_m] \in \mathbb{R}^{m\times m(1+f)}$$

i.e. C is a matrix of matrices $C_i \in \mathbb{R}^{m\times(1+f)}$ where

$$C_i = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & -\varepsilon_i \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{array}{l} \\ \\ \\ \\ \leftarrow i-\text{th row} \\ \\ \\ \\ \end{array}$$

$$D = [D_1 D_2 \ldots D_m] \in \mathbb{R}^{m\times m(1+f+p)}$$

i.e. D is a matrix of matrices $D_i \in \mathbb{R}^{m\times(1+f+p)}$ where

$$D_i = \begin{array}{c} \left[\begin{array}{cccccc} 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & 0 & & 0 \\ 0 & 0 & \cdots & \dfrac{F_i^{1+f}\theta_i^p}{M_i e_i} & & \dfrac{F_i^{p+f}\theta_i}{M_i e_i} \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{array}\right] \begin{array}{l} \\ \\ \\ \\ \leftarrow i-\text{th row} \\ \\ \\ \\ \end{array} \end{array}$$

After all and having left the problem in the most general form we have managed to create a complicated mathematical system. Of course the presentation of the model in the vector format reduces the complexity of many parameters and the respective subscripts focusing on the structure of matrices A,B,C,D. Consequently, we may conclude that the modern control theory is the only solution path for our problem. In the next sections we shall see some special cases of the general model. In order to facilitate our calculations we shall remove the weighting factor θ (equating it to unity) and restrict the averaging period p down to two years (p=2). The specific simplification does not affect the generality of our approach as the parameters θ,p do not affect the fundamental matrix A which governs the solution of the system (and the respective stability).

5.4 Special case of the general model (Model I) for (m) companies and delay factor f=1

As a first step, in order to investigate the complicated form of our problem we shall produce a special model, called Model I, by the restriction of the delay factor (f) to a specific value i.e. f=1.

The restriction above will reduce the complexity of the whole problem by reducing the rank of the vectors and matrices of section (5.3).

So substituting f=1 into the system of equations (5.3.6) the fundamental matrices A,B,C,D and the input-state-output vectors $\underline{u}_n, \underline{x}_n, \underline{y}_n$ will take the following format:

$$\underline{x}_n = \begin{bmatrix} S_{1,n} \\ S_{1,n-1} \\ S_{2,n} \\ S_{2,n-1} \\ \vdots \\ S_{m,n} \\ S_{m,n-1} \end{bmatrix} \in R^{2m}, \quad \underline{y}_n = \begin{bmatrix} P_{1,n} \\ P_{2,n} \\ \vdots \\ P_{m,n} \end{bmatrix} \in R^{m}, \quad \underline{u}_n = \begin{bmatrix} C_{1,n} \\ C_{1,n-1} \\ C_{1,n-2} \\ C_{1,n-3} \\ C_{2,n} \\ C_{2,n-1} \\ C_{2,n-2} \\ C_{2,n-3} \\ \vdots \\ C_{m,n} \\ C_{m,n-1} \\ C_{m,n-2} \\ C_{m,n-3} \end{bmatrix} \in R^{4m}$$

and $\quad A \in \mathbb{R}^{2m\times 2m}, B \in \mathbb{R}^{2m\times 4m} \qquad C \in \mathbb{R}^{m\times 2m}, D \in \mathbb{R}^{m\times 4m}$

$$A = \begin{bmatrix} R_1\lambda_{11} & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 & R_1\lambda_{31} & 0 & \cdots & R_1\lambda_{m1} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ R_2\lambda_{12} & 0 & R_2\lambda_{22} & -e_2\varepsilon_2 & R_2\lambda_{31} & 0 & \cdots & R_2\lambda_{m2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ R_m\lambda_{1m} & 0 & R_m\lambda_{2m} & 0 & R_m\lambda_{3m} & 0 & \cdots & R_m\lambda_{mm} & -e_m\varepsilon_m \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} -1 & 0 & \frac{F_1^2}{M_1'} & \frac{F_1^3}{M_1'} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & \frac{F_2^2}{M_2'} & \frac{F_2^3}{M_2'} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 0 & \frac{F_m^2}{M_m'} & \frac{F_m^3}{M_m'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C=\begin{bmatrix} 0 & -\varepsilon_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\varepsilon_m \end{bmatrix}$$

$$D=\begin{bmatrix} 0 & 0 & \frac{F_1^2}{M_1' e_1} & \frac{F_1^3}{M_1' e_1} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{F_2^2}{M_2' e_2} & \frac{F_2^3}{M_2' e_2} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{F_m^2}{M_m' e_m} & \frac{F_m^3}{M_m' e_m} \end{bmatrix}$$

where $M'_\ell = F_\ell^2 + F_\ell^3$

Before going further, we shall discuss a little more matrix A which is the most important element of the dynamic system. Actually we shall rewrite matrix A in two other forms.

$$A = R^{(R)} \cdot \Lambda_0^{(ts)} + E \text{ and } A = R^{(R)} \cdot \Lambda_1^{(ts)} + I_1$$

where the superscript (ts) stands for transpose matrix.

$$R^{(R)} = \left[r_{ij} \right] = \begin{bmatrix} R_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & R_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & R_m & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$$

so actually $$r_{ij} = \begin{cases} 0 \,, i \neq j \text{ or } i = j = 2k \,, k = 1,2,\ldots,m \\ R_{\frac{1+i}{2}} \,, i = j = 2k-1 \,, k = 1,2,\ldots,m \end{cases}$$

$$\Lambda_0 = \begin{bmatrix} \lambda_{11} & 0 & \lambda_{12} & 0 & \lambda_{13} & 0 & \cdots & \lambda_{1m} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \lambda_{21} & 0 & \lambda_{22} & 0 & \lambda_{23} & 0 & \cdots & \lambda_{2m} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{m1} & 0 & \lambda_{m2} & 0 & \lambda_{m3} & 0 & \cdots & \lambda_{mm} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} = \left[\lambda_{ij}^0\right] \in \mathbb{R}^{2m\times 2m}$$

and $\Lambda_0^{(ts)}$ is the transponse matrix of Λ_0.

As we observe Λ_0 is obtained from Λ matrix adding m-zero rows and m-zero columns. The Λ_1 matrix is obtained from Λ_0 matrix modifying some of the elements of the diagonial which lies below the first main diagonial i.e.

$$\Lambda_1 = \left[\lambda_{ij}^1\right] \in \mathbb{R}^{2m\times 2m}$$

and

$$\lambda_{ij}^1 = \begin{cases} -\dfrac{e_{\frac{i}{2}} \cdot \varepsilon_{\frac{i}{2}}}{R_{\frac{i}{2}}}, & j = 1 + i = 2k\ ,\ k = 1,2,\ldots,m \\ \lambda_{ij}^0 & ,\ \text{any other case} \end{cases}$$

$$E = \begin{bmatrix} 0 & -e_1\varepsilon_1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & -e_2\varepsilon_2 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & -e_m\varepsilon_m \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$$

and finally matrix I_1 is the same with E without the diagonial with - $e_i\varepsilon_i$ (i.e. only the one secondary diagonial non zero).

5.5 General Solution (Matrix representation) of Model I

Again, having obtained the standard format of the dynamic system we may use the standard solution method described in section (2.8). Hence,

$$\underline{x}_n = A^n \cdot \underline{x}_0 + \sum_{k=0}^{n-1} A^k \cdot B \cdot \underline{u}_{n-k-1}\ ,\ n=1,2,...$$

$$\underline{y}_n = CA^{n-1}\underline{x}_0 + CA^{n-2}B\underline{u}_0 + ... + CB\underline{u}_{n-2} + D\underline{u}_n\ ,\ \ n=1,2,...$$

Our basic concern is to examine the power series of A, (A^n, n=0,1,2,...) and obtain the diagonal (or Jordan) form of A calculating the eigenvalues and the respective eigenvectors.

$$\rho \text{ eigenvalue of } A \Leftrightarrow \varphi_m(\rho) = |\rho I - A| = 0$$

$$\varphi_m(\rho) = \begin{vmatrix} \rho - R_1\lambda_{11} & e_1\varepsilon_1 & -R_1\lambda_{21} & 0 & \cdots & -R_1\lambda_{m1} & 0 \\ -1 & \rho & 0 & 0 & \cdots & 0 & 0 \\ -R_2\lambda_{12} & 0 & \rho - R_2\lambda_{22} & e_2\varepsilon_2 & \cdots & -R_2\lambda_{m2} & 0 \\ 0 & 0 & -1 & \rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -R_m\lambda_{1m} & 0 & -R_m\lambda_{2m} & 0 & \cdots & \rho - R_m\lambda_{mm} & e_m\varepsilon_m \\ 0 & 0 & 0 & 0 & \cdots & -1 & \rho \end{vmatrix} = 0$$

The full analytical development of $\varphi_m(\rho)$ is very difficult. Here, we describe a set of recursive relationships which help us to calculate the required $\varphi_m(\rho)$.

Firstly, we define the symbol $M_{i_1 i_2 \ldots i_k}$ which represents the minor determinant which arises from $\varphi_m(\rho)$ deleting the i_1-th, i_2-th,...,i_k-th rows and columns. (Trivially $M_0 = \varphi_m(\rho)$).

Developing across the second row we obtain

$\varphi_m(\rho) = e_1 \cdot \varepsilon_1 \cdot M_{12} + \rho \cdot M_2$ $\quad$ 1^{st}-Level

---

$M_{12} = e_2\varepsilon_2 \cdot M_{1234} + \rho \cdot M_{124}$

$\quad$ 2^{nd}-Level

$M_2 = e_2\varepsilon_2 \cdot M_{234} + \rho \cdot M_{24}$

$M_{1234} = e_3\varepsilon_3 \cdot M_{123456} + \rho \cdot M_{12346}$

$M_{124} = e_3\varepsilon_3 \cdot M_{12456} + \rho \cdot M_{1246}$

$M_{234} = e_3\varepsilon_3 \cdot M_{23456} + \rho \cdot M_{2346}$ $\quad$ 3^{rd}-Level

$M_{24} = e_3\varepsilon_3 \cdot M_{2456} + \rho \cdot M_{246}$

...

We systematically develop the determinants across the first row (from the top) which has the elements of (-1) and ρ, and finally obtain a polynomial of (2m)-degree i.e.

$$\varphi_m(\rho) = a_{2m}\rho^{2m} + a_{2m-1}\rho^{2m-1} + \ldots + a_1\rho + a_0$$

As regards the coefficients of $\varphi_m(\rho)$ we may observe that generally

$$a_k = f(e_i, \varepsilon_i, \lambda_{ij}), \ k=1,2,\ldots,2m$$

From the form of $\varphi_m(\rho)$ and the development we have described before we may derive that

$$a_{2m} = 1$$

(As all the terms containing ρ are lying on the diagonios of $\varphi_m(\rho)$ with no coefficient)

$$a_0 = e_1 e_2 \ldots e_m \varepsilon_1 \varepsilon_2 \ldots \varepsilon_m$$

(if we consider the first row of each development level described before, we can see that the constant term will be derived as $a_0 = e_1\varepsilon_1 \cdot e_2\varepsilon_2 \cdot \ldots \cdot e_m\varepsilon_m \cdot M_{12\ldots2m}$ where $M_{12\ldots2m} = 1$).

The full determination of the roots of $\varphi_m(\rho)=0$ (eigenvalues of A) requires numerical methods. Hence, after calculating the eigenvalues and the respective eigenvectors we may calculate the power series of A^n (and consequently the solution). Another approach for the calculation of A^n may use the analysis for matrix A of section (5.4).

5.6 Stability Analysis of Model I

In this section we shall examine the stability of the model I (see analysis of section (2.15)).

First of all we must calculate the equilibrium points $\underline{x}_e$

$$\underline{x}_e \text{ equilibrium point} \Leftrightarrow A\underline{x}_e = \underline{x}_e$$

As we can see,

$$\det(A)=\begin{vmatrix} R_1\lambda_{11} & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 & \cdots & R_1\lambda_{m1} & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ R_2\lambda_{12} & 0 & R_2\lambda_{22} & -e_2\varepsilon_2 & \cdots & R_2\lambda_{m2} & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ R_m\lambda_{1m} & 0 & R_m\lambda_{2m} & 0 & \cdots & R_m\lambda_{mm} & -e_m\varepsilon_m \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{vmatrix}=$$

(developing across the second row and then, across the first column we obtain)

$$\det(A)=(-1)(-e_1\varepsilon_1)\begin{vmatrix} R_2\lambda_{22} & -e_2\varepsilon_2 & \cdots & R_2\lambda_{m2} & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & R_2\lambda_{22} & \vdots & \vdots \\ R_m\lambda_{1m} & 0 & & R_m\lambda_{mm} & -e_m\varepsilon_m \\ 0 & 0 & \cdots & 1 & 0 \end{vmatrix}$$

So, there is a recursive relationship. If we define

$$\varphi(e_1,e_2,e_3,e_m,\varepsilon_1,\varepsilon_2,\varepsilon_3,\ldots,\varepsilon_m)=\det(A)$$

then, $\det(A)=e_1\varepsilon_1\varphi(e_2,e_3,\ldots,e_m,\varepsilon_2,\varepsilon_3,\ldots,\varepsilon_m)=$

$$=e_1\varepsilon_1\varepsilon_2\varepsilon_2\varphi(\varepsilon_3,\ldots,e_m,\varepsilon_3,\ldots,\varepsilon_m)=$$

$$\ldots$$

$$=e_1\varepsilon_1 e_2\varepsilon_2\ldots e_{m-1}\cdot\varepsilon_{m-1}\cdot\varphi(e_m,\varepsilon_m)$$

$$\varphi(e_m,\varepsilon_m)=\begin{vmatrix} R_m\lambda_{mm} & -e_m\varepsilon_m \\ 1 & 0 \end{vmatrix}=e_m\varepsilon_m$$

Hence, $\det(A)=e_1e_2\cdot\ldots\cdot e_m\cdot\varepsilon_1\varepsilon_2\cdot\ldots\cdot\varepsilon_m \neq 0$ (given that $e_i\neq 0$ and $\varepsilon_i\neq 0$ $\forall i$) and consequently the homogeneous system has only the zero solution which is the only equilibrium point in the state space.

The Liapunov or asymptotic stability should be considered only at the zero point $\underline{0}$.

The examination of system's stability is very difficult as we have not analytic formulae for all the coefficients of the characteristics polynomial and consequently for the eigenvalues of matrix A.

The only possible action for the full investigation of the system's stability is the determination of the eigenvalues of A with numerical methods for different sets of the parameter values and then to see whether the eigenvalues lie in the unit circle.

Before starting this hard work, we may apply an easy criterion for instability which is easily derived from the analytic form of the first and last coefficients a_0 and a_{2m} of the $\varphi_m(\rho)$.

As we know, $a_{2m}=1$ and $a_0=e_1e_2...e_m\varepsilon_1\varepsilon_2...\varepsilon_m$. From the theory of polynomials we obtain that the product of the possible roots of the equation $\varphi_m(\rho)=0$ equals the ratio of the last and first coefficient.

If $\rho_1,\rho_2,...,\rho_{2m}$ are the roots then

$$|\rho_1 \cdot \rho_2 \cdot ... \cdot \rho_{2m}| = e_1 e_2 \cdot ... \cdot e_m \cdot \varepsilon_1 \varepsilon_2 \cdot ... \cdot \varepsilon_m$$

Now if the last product of the expense and profit vector s is greater than unity then there is always a root with absolute value greater than unity and the system will be unstable.

So, designing the system, we should choose the expense and profit vectors such that the last product is less than the unity.

Closing the section, we should mention that the transfer matrix function may be also obtained using the standard formula

$$G(2) = C[zI - A]^{-1} B + D$$

but no further insight is gained and of course this requires a lot of algebra work.

5.7 Controllability & Observability properties of Model I

As we have seen in sections (2.13) & (2.14) there are two properties of the system known as controllability and observability which determine the existence of an optimal control.

Another basic result states that the existence of the two properties allows us the ability to regulate the output which of course is one of our major concerns.

So in this section we shall investigate the two properties mentioned above.

Controllability

We have to examine the rank of $\mathscr{B}$ matrix and prove that rank $\mathscr{B}$=2m for complete controllability of the system (see theorem (2.13.1) i.e.

$$\text{rank}\mathscr{B} = \text{rank}\left[B \vdots AB \vdots A^2B \vdots \cdots \vdots A^{2m-1}B\right] = 2m$$

As we can see $\mathscr{B} \in \mathbb{R}^{2m \times 8m^2}$, so there is a lot of calculation work. We shall avoid it, using a trick (having the B matrix and calculating only the AB product).

Matrix $\mathscr{B}$ is written $[\underline{b}_1 \underline{b}_2 \ldots \underline{b}_{4m} \underline{b}_{4m+1} \ldots \underline{b}_{8m} \underline{b}_{8m+1} \ldots \underline{b}_{8m^2}]$. The first 4m vectors belong to matrix B and as we can see the $\underline{b}_1, \underline{b}_5, \underline{b}_9, \ldots, \underline{b}_{4m-3}$ are linearly independent (i.e. m vectors). We should find another m linear independent vectors in order to fulfill the requirement of 2m.

Let us calculate the product AB.

$$AB = \begin{bmatrix} -R_1\lambda_{11} & 0 & \frac{F_1^2}{M_1'}R_1\lambda_{11} & \frac{F_1^3}{M_1'}R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \frac{F_2^2}{M_2'}R_1\lambda_{21} & \frac{F_2^3}{M_2'}R_1\lambda_{21} & \cdots \\ -1 & 0 & \frac{F_1^2}{M_1'} & \frac{F_1^3}{M_1'} & 0 & 0 & 0 & 0 & \cdots \\ -R_2\lambda_{22} & 0 & \frac{F_1^2}{M_1'}R_2\lambda_{12} & \frac{F_1^3}{M_1'}R_2\lambda_{12} & -R_2\lambda_{22} & 0 & \frac{F_2^2}{M_2'}R_2\lambda_{22} & \frac{F_2^3}{M_2'}R_2\lambda_{12} & \cdots \\ 0 & 0 & 0 & 0 & -1 & 0 & \frac{F_2^2}{M_2'} & \frac{F_2^3}{M_2'} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

and we can rewrite $AB = [\underline{b}_{4m+1} \underline{b}_{4m+2} \ldots \underline{b}_{8m}]$.

The m vectors $\underline{b}_{4m+1}, \underline{b}_{4m+5}, \ldots, \underline{b}_{8m-3}$ are linearly independent and all the vectors

$$\underline{b}_1, \underline{b}_5, \ldots, \underline{b}_{4m-3}, \underline{b}_{4m+1}, \underline{b}_{4m+5}, \ldots, \underline{b}_{8m-3}$$

are linearly independent.

Hence rank $\mathscr{B}$=2m and consequently the system is completely controllable.

Observability

From theorem (2.14.1) we obtain that the system (A,B,C,D) is complete observable if and only if

$$\operatorname{rank}\mathscr{C} = \operatorname{rank}\begin{bmatrix} C \\ \cdots \\ CA \\ \cdots \\ \vdots \\ \cdots \\ CA^{2m-1} \end{bmatrix} = 2m$$

where $\mathscr{C} \in \mathbb{R}^{2m^2 \times 2m}$.

We shall use the same trick as for the controllability property. The $\mathscr{C}$ matrix may be written in the form

$$\mathscr{C} = \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \vdots \\ \underline{c}_{2m^2} \end{bmatrix}$$

where $$C = \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \vdots \\ \underline{c}_m \end{bmatrix}, CA = \begin{bmatrix} \underline{c}_{m+1} \\ \underline{c}_{m+2} \\ \vdots \\ \underline{c}_{2m} \end{bmatrix}, CA^{2m-1} = \begin{bmatrix} \underline{c}_{2m^2-m+1} \\ \underline{c}_{2m^2-m+1} \\ \vdots \\ \underline{c}_{2m^2} \end{bmatrix}$$

Again we observe that $\underline{c}_1, \underline{c}_3, \ldots, \underline{c}_{m-1}$ vectors are linear independent. We shall calculate the product C·A.

$$CA = \begin{bmatrix} -\varepsilon_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\varepsilon_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\varepsilon_m & 0 \end{bmatrix}$$

Obviously, $\underline{c}_{m+1}, \underline{c}_{m+2}, \ldots, \underline{c}_{2m}$ are linear independent and all of them $\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_{2m}$ are linearly independent and consequently

$$\operatorname{rank} \mathscr{C} = 2m$$

Hence, the system is completely observable.

5.8 Optimal design for the matrices of Model I

The optimal design for the parameter values involved in the problem is very interesting but it is very difficult to have an analytical approach. In this section we shall try to describe a method which may provide the solution with the fastest response to different input signals. For this purpose we shall formalize the procedure.

Let $s = (\underline{e}, \underline{\varepsilon}, \underline{R}, \Lambda)$ a choice for the parameter values and S^* the whole set of choices for s. Let $\rho_s = \max\{|\rho_{s,1}|, |\rho_{s,2}|, \ldots, |\rho_{s,2m}|\}$ be the maximum absolute value of the eigenvalues of matrix A with respect to s. Actually ρ_s is the radius of a circle with its center placed on zero which contains all the eigenvalues $\rho_{s,1}, \rho_{s,2}, \ldots, \rho_{s,2m}$.

As the solution of the system is a linear combination of the power series of the eigenvalues then the magnitude of the solution depends on the maximum absolute value of the eigenvalue of A (i.e. it depends on ρ_s).

Hence, the fastest response is obtained for the choice s which minimizes the ρ_s.

Let,

$$\rho_0 = \min\{\rho_S : s \in S^*\}$$

We shall show that,

$$\rho_0 \geq \sqrt[2m]{e_1 e_2 \cdot \ldots \cdot e_m \varepsilon_1 \varepsilon_2 \cdot \ldots \cdot \varepsilon_m} \tag{5.8.1}$$

As we have shown in section (5.5) the characteristic polynomial has the following form

$$\varphi_m(\rho) = a_{2m}\rho^{2m} + a_{2m-1}\rho^{2m-1} + \ldots + a_1\rho + a_0$$

with $a_{2m}=1$ and $a_0=e_1\cdot\ldots\cdot e_m\cdot\varepsilon_1\cdot\ldots\cdot\varepsilon_m$

Consequently, using the argument of section (5.6)

$$\rho_s^{2m} \geq |\rho_{s,1}|\cdot|\rho_{s,2}|\cdot\ldots\cdot|\rho_{s,2m}| = e_1\cdot\ldots\cdot e_m\cdot\varepsilon_1\cdot\ldots\cdot\varepsilon_m$$

From the last inequality we obtain (5.8.1)

The minimization of the maximum root of $\varphi_m(\rho)$ is obtained in two cases.

1st case: When the $\varphi_m(\rho)$ has a real root with multiplicity equal to the degree of the polynomial (i.e. 2m).

Then $$\varphi_m(\rho) = (\rho - \rho_0)^{2m} \tag{5.8.2}$$

where $$\rho_0 = \sqrt[2m]{e_1\cdot\ldots\cdot e_m\cdot\varepsilon_1\cdot\ldots\cdot\varepsilon_m} \tag{5.8.3}$$

Assuming there is the ρ_0 we may expand the (5.8.2) and obtain

$$\varphi_m(\rho) = (\rho - \rho_0)^{2m} = \sum_{k=0}^{2m} \binom{2m}{k}\cdot(-\rho_0)^{2m-k}\cdot\rho^k \tag{5.8.4}$$

then equating the respective coefficients

$$a_k = \binom{2m}{k}(-\rho_0)^{2m-k},\ k=0,1,\ldots,2m \tag{5.8.5}$$

we obtain a system of (2m+1) equations which contain m^2+3m (control) parameters, i.e. $e_1,\ldots,e_m$, $\varepsilon_1,\ldots,\varepsilon_m$, $R_1,\ldots,R_m$, $\lambda_{11},\ldots,\lambda_{mm}$. Some of them may be fully ($\underline{\varepsilon}$ vector and Λ matrix) or partially ($\underline{e}$ and $\underline{R}$ vectors) controlled. Our aim should be the optimal selection of all the controlled parameters such that the system becomes solveable. In that case we may obtain a root (for the characteristic polynomial) with multiplicity 2m.

2nd case: When the $\varphi_m(\rho)$ has roots (real or complex) which all lie on the circumference of the circle with the origin at zero and radius $\rho_0 = \sqrt[2m]{e_1 \cdot \ldots \cdot e_m \cdot \varepsilon_1 \cdot \ldots \cdot \varepsilon_m}$.

i.e. $\rho_j = \rho_0 \cdot (\cos\theta_j + i\sin\theta_j)$ j=1,2,...,m

and the polynomial $\varphi_m(\rho)$ will have the form

$$\varphi_m(\rho) = \rho^{2m} + e_1 \cdot \ldots \cdot e_m \cdot \varepsilon_1 \cdot \ldots \cdot \varepsilon_m \tag{5.8.6}$$

Hence the second case may appears when

$$a_{2m-1} = a_{2m-2} = \ldots = a_2 = a_1 = 0 \tag{5.8.7}$$

(i.e. all the coefficients except the very first and last one should be equal to zero).

Obviously, for large m the systems of equations becomes very complicated or there is not a choice of s resulting a root with multiplicity 2m. That's why, we should be forced to follow a "trial and error" procedure with numerical methods with the following steps.

Step 1: Make an initial guess for the parameter values (which are controllable) and obtain the choise s_1 and the respective ρ_{s_1} (after solving the polynomial $\varphi_m(\rho)$ with numerical methods).

Step 2: Compare the magnitude of ρ_{s_1} with the ρ_0 (always $\rho_0 < \rho_s^1$) and then try to change one by one the controllable parameters obtaining a choice s_2 and ρ_{s_2}.

Step 3: Compare ρ_{s_1}, ρ_{s_2} and ρ_0 and try to identify the effect of the change of each parameter. Then, again a choose s_3 and calculate ρ_{s_3}.

...

Of course we need a lot of simulation work with a computer and after (sp) steps decide whether

(a) ρ_{s_n} may converge to ρ_0 and so obtain a choice s^* and $\rho_{s^*}:\left|\rho_0-\rho_{s^*}\right|<\delta$

(δ is the desired distance) continuing the simulation.

(b) ρ_{s^*} may not converge to ρ_0 and so obtain the best choice s^{**} of our simulaton which gives the $\rho_{s^{**}}$ such that

$$\left|\rho_0-\rho_{s^{**}}\right|=\min\left\{\left|\rho_0-\rho_{s_n}\right|:n=1,2,\ldots,(sp)\right\}$$

5.9 Special case of the general model (Model II) for (m=2) companies and (f) delay factor

In this section, we shall provide another version (Model II) of our general model restricting the first parameter i.e. the number of participating companies m=2 while leaving as a free parameter the delay factor (f).

So again substituting in thc system of equation (5.3.6) the value of m=2 we obtain the input - state - output vectors ($\underline{u}_n,\underline{x}_n,\underline{y}_n$) respectively and the fundamental matrices A,B,C,D.

$$\underline{X}_n = \begin{bmatrix} S_{1,n} \\ S_{1,n-1} \\ \vdots \\ S_{1,n-f} \\ S_{2,n} \\ S_{2,n-1} \\ \vdots \\ S_{2,n-f} \end{bmatrix} \in \mathbb{R}^{2(1+f)}, \quad \underline{y}_n = \begin{bmatrix} P_{1,n} \\ P_{2,n} \end{bmatrix} \in \mathbb{R}^{2}, \quad \underline{u}_n = \begin{bmatrix} C_{1,n} \\ C_{1,n-1} \\ \vdots \\ C_{1,n-f-1} \\ C_{1,n-f-2} \\ C_{2,n} \\ C_{2,n-1} \\ \vdots \\ C_{2,n-2-1} \\ C_{2,n-f-2} \end{bmatrix} \in \mathbb{R}^{2(3+f)}$$

and $\quad A \in \mathbb{R}^{2(1+f)\times 2(1+f)}$, $B \in \mathbb{R}^{2(1+f)\times 2(3+f)}$

$C \in \mathbb{R}^{2\times 2(1+f)}$, $D \in \mathbb{R}^{2\times 2(3+f)}$

$$A = \left[\begin{array}{cccccc:cccccc}
R_1\lambda_{11} & 0 & 0 & \cdots & 0 & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 & 0 & \cdots & 0 & 0 \\
1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\hdashline
R_2\lambda_{12} & 0 & 0 & \cdots & 0 & 0 & R_2\lambda_{22} & 0 & 0 & \cdots & 0 & -e_2\varepsilon_2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0
\end{array}\right]$$

$$B = \left[\begin{array}{cccccc:cccccc}
-1 & 0 & \cdots & 0 & \frac{F_1^2}{M_1'} & \frac{F_1^3}{M_1'} & & & & \mathbb{O} & & \\
 & & \mathbb{O} & & & & & & & & & \\
\hdashline
 & & & & & & -1 & 0 & \cdots & 0 & \frac{F_2^2}{M_2'} & \frac{F_2^3}{M_2'} \\
 & & \mathbb{O} & & & & & & & \mathbb{O} & &
\end{array}\right]$$

$$C=\left[\begin{array}{ccccc|ccccc} 0 & 0 & \cdots & 0 & -\varepsilon_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -\varepsilon_2 \end{array}\right]$$

$$D=\left[\begin{array}{cccccc|cccccc} 0 & 0 & \cdots & 0 & \frac{F_1^2}{M_1'e_1} & \frac{F_1^3}{M_1'e_1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \frac{F_2^2}{M_2'e_2} & \frac{F_2^2}{M_2'e_2} \end{array}\right]$$

5.10 General Solution of Model II

Again having obtained the standard format of the dynamic system we may use the standard solution method described in section (2.8) i.e. equations (2.8.1) and (2.8.2).

Similarly with the work of section (5.5), we should proceed with the calculation of the power series of A (A^n, n=1,2,...), so actually we have to obtain the eigenvalues and the respective eigenvectors of matrix A. Hence,

$$\rho \text{ eigenvalue of } A \Leftrightarrow \varphi_2(\rho)=|\rho I-A|=0 \text{ or}$$

$$\varphi_2(\rho)=\left[\begin{array}{ccccc|ccccc} \rho-R_1\lambda_{11} & 0 & \cdots & 0 & e_1\varepsilon_1 & -R_1\lambda_{21} & 0 & \cdots & 0 & 0 \\ -1 & \rho & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & \rho & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & -1 & \rho & 0 & 0 & \cdots & 0 & 0 \\ \hline -R_2\lambda_{12} & 0 & \cdots & 0 & 0 & \rho-R_2\lambda_{22} & 0 & \cdots & 0 & e_2\varepsilon_2 \\ 0 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & \rho & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & \rho \end{array}\right]=0$$

The analytical development of $\varphi_2(\rho)$ is again difficult but we **can** find the final form if we follow a simple rule and a strict recursive procedure.

Firstly, the simple rule: Always develop the major determinant $\varphi_2(\rho)$ or the minor ones (which are produced deleting rows and columns) across the (first) row or column which has the greatest number of zero elements.

Secondly we shall describe the recursive procedure with the following steps.

Step 1: Develop the $\varphi_2(\rho)$ across the second row which has only two non zero elements the -1 and ρ and so,

$$\varphi_2(\rho) = (-1)\cdot(-1)\cdot \Xi_1^2 + \rho \cdot \Psi_1^2 \tag{5.10.1}$$

where Ξ_1^2 is the minor determinant of $\varphi_2(\rho)$ which is produced, deleting the 1st column and the 2nd row of $\varphi_2(\rho)$ and

Ψ_1^2 is the minor determinant of $\varphi_2(\rho)$ which is produced, deleting the 2nd column and the 2nd row of $\varphi_2(\rho)$.

[The (2) superscript of Ξ_1 and Ψ_1 is referred to $\varphi_2(\rho)$]

Step 2: Develop the minor determinant Ξ_1^2 across the first column (which has only one non-zero element the -1). So,

$$\Xi_1^2 = (-1)\cdot(-1)\cdot \Xi_2^2 \tag{5.10.2}$$

Step 3: Continue to develop the

$$\Xi_i^2 = (-1)(-1)\Xi_{i+1}^2, \quad i=2,3,\ldots,f-1 \tag{5.10.3}$$

where Ξ_i (i=2,3,...,f) is the minor determinant of Ξ_{i-1} which is produced, deleting the 1st column and 2nd row of Ξ_i and

$$\Xi_f^2 = \begin{vmatrix} e_1\varepsilon_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boxed{\varphi_1(\rho)} & \\ 0 & & & \end{vmatrix} \tag{5.10.4}$$

From the last equation we obtain

$$\Xi_f^2 = e_1\varepsilon_1 \cdot \varphi_1(\rho) \tag{5.10.5}$$

Step 4: Combine equations (5.10.1), (5.10.2), (5.10.3) and (5.10.5) and obtain that

$$\varphi_2(\rho) = e_1 e_1 \cdot \varphi_1(\rho) + \rho \cdot \Psi_1^2 \tag{5.10.6}$$

Step 5: Develop the minor determininant Ψ_1 across the second row (which has only one non-zero element, ρ) i.e.

$$\Psi_1^2 = \rho \cdot \Psi_2^2 \tag{5.10.7}$$

Step 6: Continue to develop Ψ_i, i=2,3,..., determinants across the second row (as with Ψ_1) and finally obtain,

$$\Psi_i^2 = \rho \cdot \Psi_{i+1}^2, \quad i=2,3,\ldots,f-1 \tag{5.10.8}$$

where Ψ_i^2 (i=2,3,...,f) is the minor determinant of Ψ_{i-1}^2 which is produced, deleting the 2[nd] column and the second row and

$$\Psi_f^2 = \begin{vmatrix} \rho - R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \cdots & 0 \\ -R_2\lambda_{12} & & & & \\ 0 & & \varphi_1(\rho) & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix} \tag{5.10.9}$$

Step 7: Combining equations (5.10.6), (5.10.7) and (5.10.8) we obtain

$$\varphi_2(\rho) = e_1\varepsilon_1 \cdot \varphi_1(\rho) + \rho^f \cdot \Psi_f^2 \tag{5.10.10}$$

Step 8: Develop Ψ_f^2 across the third row which has two non-zero elements -1 and ρ. So we obtain,

$$\Psi_f^2=(-1)^2\cdot\begin{vmatrix}\rho-R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \cdots & 0\\ -R_2\lambda_{12} & & & & \\ 0 & & \boxed{\quad\Xi_1^1\quad} & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix}+\rho\cdot\begin{vmatrix}\rho-R_1\lambda_{11} & -R_1\lambda_{21} & 0 & \cdots & 0\\ -R_2\lambda_{12} & & & & \\ 0 & & \boxed{\quad\Psi_1^1\quad} & & \\ \vdots & & & & \\ 0 & & & & \end{vmatrix}$$

(now the (1) superscripts of Ξ_1 and Ψ_1 refers to $\varphi_1(\rho)$). Ξ_1^1 and Ψ_1^1 are produced similarly to Ξ_1^2 and Ψ_1^2 from $\varphi_n(\rho)$. So we follow similar steps and finally,

$$\Psi_f^2=\begin{vmatrix}\rho-R_1\lambda_{11} & 0 & 0\\ -R_2\lambda_{12} & 0 & e_2\varepsilon_2\\ 0 & -1 & \rho\end{vmatrix}+p^{f-1}\begin{vmatrix}\rho-R_1\lambda_{11} & -R_1\lambda_{21} & 0\\ -R_2\lambda_{12} & \rho-P_2L_{22} & e_2\varepsilon_2\\ 0 & 0 & \rho\end{vmatrix}\Rightarrow$$

$$\Psi_f^2=e_2\varepsilon_2(\rho-R_1\lambda_{11})+\rho^f\left[(\rho-R_1\lambda_{11})(\rho-R_2\lambda_{22})-R_1R_2\lambda_{12}\lambda_{21}\right] \tag{5.10.11}$$

Step 9: Develop $\varphi_1(\rho)$ similarly with $\varphi_2(\rho)$ and obtain

$$\varphi_1(\rho)=e_2\varepsilon_2+\rho^f(\rho-R_2\lambda_{22}) \tag{5.10.12}$$

Step 10: Combine equations (5.10.10), (5.10.11) and (5.10.12) and finally obtain

$$\varphi_2(\rho)=e_1e_2\varepsilon_1\varepsilon_2+e_1\varepsilon_1\rho^f(\rho-R_2\lambda_{22})+\rho^f e_2\varepsilon_2(\rho-R_1\lambda_{11})$$

$$+\rho^{2f}\cdot(\rho-R_1\lambda_{11})(\rho-R_2\lambda_{22})-\rho^{2f}R_1R_2\lambda_{12}\lambda_{21} \tag{5.10.13}$$

Again we observe that $\varphi_2(\rho)$ is a polynomial with $a_{2(1+f)}=1$ (coefficient of $\rho^{2(1+f)}$ and $a_0=e_1e_2\varepsilon_1\varepsilon_2$ constant term. The full determination of the roots of $\varphi_2(p)=0$ (eigenvalues of A) requires numerical methods. After obtaining those eigenvalues and the respective eigenvectors we may calculate the solution using the standard equations.

5.11 Stability Analysis of Model II

In this sections, we shall examine the stability of model II similarly with section (5.6). Firstly, we calculate the equilibrium points $\underline{x}_e$.

$\underline{x}_e$ equilibrium point $\Leftrightarrow A \cdot \underline{x}_e = \underline{x}_e$

The equation above has at least the zero solution. We shall show that this is the only one (unique)

$$\det(A)=\left|\begin{array}{ccccc|ccccc} R_1\lambda_{11} & 0 & \cdots & 0 & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & & & & & \\ \vdots & \vdots & & \vdots & \vdots & & & \mathbb{O} & & \\ 0 & 0 & \cdots & 0 & 0 & & & & & \\ 0 & 0 & \cdots & 1 & 0 & & & & & \\ \hline R_2\lambda_{12} & 0 & \cdots & 0 & 0 & R_2\lambda_{22} & 0 & \cdots & 0 & -e_2\varepsilon_2 \\ & & & & & 1 & 0 & \cdots & 0 & 0 \\ & & \mathbb{O} & & & \vdots & \vdots & & \vdots & \vdots \\ & & & & & 0 & 0 & \cdots & 0 & 0 \\ & & & & & 0 & 0 & \cdots & 1 & 0 \end{array}\right| = \left|\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right|$$

where $A_{11}, A_{12}, A_{21}, A_{22} \in \mathbb{R}^{(1+f)\times(1+f)}$.

Developing det(A) across the 2nd row and again the minor across the 2nd row... continuously after f-steps we obtain

$$\det(A)=(-1)^f \cdot \left|\begin{array}{ccccc} -e_1\varepsilon_1 & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ 0 & & \boxed{A_{22}} & & \\ \vdots & & & & \\ 0 & & & & \end{array}\right|$$

Then developing across the 1st-row we obtain

$$\det(A)=(-1)^{f+1}e_1\varepsilon_1 \det(A_{22}) \tag{5.11.1}$$

Similarly, we develop $\det(A_{22})$ and combining also equation (5.11.1) we derive that

$$\det(A)=(-1)^{2(1+f)}e_1e_2\varepsilon_1\varepsilon_2 = e_1e_2\varepsilon_1\varepsilon_2 \tag{5.11.2}$$

From equation (5.11.2) and for $e_1,e_2,\varepsilon_1,\varepsilon_2 \neq 0$ we obtain det(A) i.e. $\underline{0}$ is the only solution to the system $A \cdot \underline{x}_e = \underline{x}_e$ consequently the only equilibrium point.

Now although we have the analytical form of the $\varphi_2(\rho)$ characteristic polynomial it is very difficult to investigate and find the position of the eigenvalues of matrix A.

Again numerical methods and arguments as in section (5.6) should be applied in order to examine the magnitude of the eigenvalues (less or greater than unity). Of course we can also use the root-locus method to investigate the potential stability (against a certain parameter). For example if we want to examine the root loci against the values of parameter ε_1 we may rearrange equation (5.10.13) in the form of Step 1 of Appendix III i.e.

$$\varphi_2(\rho)=1+\varepsilon_1 \frac{e_1 e_2 \varepsilon_2 + e_1 \rho^f (p - R_2\lambda_{22})}{\rho^{2f}(\rho - R_1\lambda_{11})(\rho - R_2\lambda_{22}) - \rho^{2f} R_1 R_2 \lambda_{12}\lambda_{21} + \rho^f e_2 \varepsilon_2 (\rho - R_1\lambda_{11})} \qquad (5.11.3)$$

The procedure described in Appendix III requires the zeros and poles of the fraction which appears in equation (5.11.3). Again these values can only be obtained using numerical methods (for a large f value).

5.12 Controllability and Observability properties of Model II

In this section we shall examine the concepts of controllability and observability of model II.

Controllability

Similarly, with section (5.7) we must prove

$$\operatorname{rank}\mathscr{B} = \left[B \vdots AB \vdots A^2B \vdots \cdots \vdots A^{2f+1}B \right] = 2(1+f)$$

for the complete controllability of the system.

We shall show that taking the first column of each of the 2(1+f) matrices included in $\mathscr{B}$ we can obtain 2(1+f) linear independent vectors and consequently

$$\text{rank} \quad \mathscr{B}=2(1+f).$$

We may observe that matrix B has only one non-zero element in its first column (at the top, equal to -1). If we calculate consecutively AB, $A^2B,\ldots,A^{2f+1}B$ we may observe that this non-zero element (-1) is going down one by one row in the first column. So we may actually obtain the 2(1+f) linear independent vector. So model II is completely controllable.

Observability

For the observability property we should prove that

$$\text{rank}\,\mathscr{C}=\begin{bmatrix} C \\ \cdots \\ CA \\ \cdots \\ \vdots \\ \cdots \\ CA^{2f+1} \end{bmatrix}=2(1+f)$$

Calculating carefully the required products we obtain

$$\mathscr{C}=\left[\begin{array}{ccccccc|cccccc}
0 & 0 & 0 & \cdots & 0 & 0 & -\varepsilon_1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -\varepsilon_2 \\
\hline
0 & 0 & 0 & \cdots & 0 & -\varepsilon_1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -\varepsilon_2 & 0 \\
\hline
0 & 0 & 0 & \cdots & -\varepsilon_1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -\varepsilon_2 & 0 & 0 \\
\hline
 & & & & & & \vdots & & \vdots & & & & &
\end{array}\right]$$

As we observe the $\varepsilon_1,\varepsilon_2$ are moving to the left (creating new linear independent vector) and easily identify 2(1+f) linear indepent vectors, so rank $\mathscr{C}$=2(1+f) and the system is completely observable.

5.13 Optimal design for the matrices of Model II

The discussion is exactly the same as for the Model I in Section (5.8). In this case we have a characteristic polynomial of 2(1+f) degree see equation (5.10.13) with

$$a_{2(1+f)} = 1 \quad \text{and} \quad a_0 = e_1 e_2 \varepsilon_1 \varepsilon_2$$

So according to inequality (5.8.1)

$$\rho_0 \geq \sqrt[2(1+f)]{e_1 e_2 \varepsilon_1 \varepsilon_2}$$

We may explore as in section (5.8) the optimal set of parameter values for $\underline{e}, \underline{\varepsilon}, \underline{R}$ vectors and Λ matrix using the algorithm described there in order to obtain (if possible) a root with 2(1+f) multiplicity or 2(1+f) roots lying in the circumference of the circle with center at zero and radius $\sqrt[2(1+f)]{e_1 e_2 \varepsilon_1 \varepsilon_2}$ for the characteristic polynomial and consequently the fastest output response for the system.

5.14 Special cases of the general model (Models III, IV) for (m=2) companies and delay factor (f=1) with numerical examples

In this section we shall firstly examine a very special case restricting the number of participating companies m=2 and the delay factor f=1 (Model III). Actually this third Model may be described as the intersection point of Models I and II.

Hence, we obtain the following system of equation (which corresponds to the general system (5.3.6) under the modifications and restrictions mentioned above.

$$\left.\begin{aligned}
S_{1,n} &= R_1\lambda_{11}S_{1,n-1} + R_1\lambda_{21}S_{2,n-1} + \frac{F_1^3}{M_1'}C_{1,n-3} + \frac{F_1^2}{M_1'}C_{1,n-2} - e_1\varepsilon_1 S_{1,n-2} - C_{1,n} \\
S_{2,n} &= R_2\lambda_{12}S_{1,n-1} + R_2\lambda_{22}S_{2,n-1} + \frac{F_2^3}{M_2'}C_{2,n-3} + \frac{F_2^2}{M_2'}C_{2,n-2} - e_2\varepsilon_2 S_{2,n-2} - C_{2,n} \\
P_{1,n} &= \frac{F_1^3}{M_1'e_1}C_{1,n-3} + \frac{F_1^2}{M_1'e_1}C_{1,n-2} - \varepsilon_1 S_{1,n-2} \\
P_{2,n} &= \frac{F_2^3}{M_2'e_2}C_{2,n-3} + \frac{F_2^2}{M_2'e_2}C_{2,n-2} - \varepsilon_2 S_{2,n-2}
\end{aligned}\right\} \qquad (5.14.1)$$

Again having obtained the system of equation we may pass to the standard dynamic format system of (5.3.7). Consequently we obtain the following vectors and matrices which again correspond to vector $\underline{x}_n, \underline{y}_n, \underline{u}_n$ and matrices A,B,C,D of section (5.3).

$$\underline{x}_n = \begin{bmatrix} S_{1,n} \\ S_{1,n-1} \\ S_{2,n} \\ S_{1,n-1} \end{bmatrix} \in \mathbb{R}^4 \quad , \quad \underline{y}_n = \begin{bmatrix} P_{1,n} \\ P_{2,n} \end{bmatrix} \in \mathbb{R}^2 \quad , \quad \underline{u}_n = \begin{bmatrix} C_{1,n} \\ C_{1,n-1} \\ C_{1,n-2} \\ C_{1,n-3} \\ C_{2,n} \\ C_{2,n-1} \\ C_{2,n-2} \\ C_{2,n-3} \end{bmatrix} \in \mathbb{R}^8$$

$$A \in \mathbb{R}^{4\times4} \; , B \in \mathbb{R}^{4\times8} \; , C \in \mathbb{R}^{2\times4} \; , D \in \mathbb{R}^{2\times8}$$

$$A = \begin{bmatrix} R_1\lambda_{11} & -e_1\varepsilon_1 & R_1\lambda_{21} & 0 \\ 1 & 0 & 0 & 0 \\ R_2\lambda_{12} & 0 & R_2\lambda_{22} & -e_2\varepsilon_2 \\ 0 & 0 & 1 & 0 \end{bmatrix} , \; D = \begin{bmatrix} 0 & 0 & \frac{F^2}{M_1'e_1} & \frac{F^3}{M_1'e_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{F^2}{M_2'e_2} & \frac{F^3}{M_2'e_2} \end{bmatrix}$$

$$B=\begin{bmatrix} -1 & 0 & \dfrac{F_1^2}{M_1'} & \dfrac{F_1^3}{M_1'} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & \dfrac{F_2^2}{M_2'} & \dfrac{F_2^3}{M_2'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad C=\begin{bmatrix} 0 & -\varepsilon_1 & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_2 \end{bmatrix}$$

Closing this section, we shall present a diagram incorporating the input-state-output concepts of the dynamic system. The diagram is not formal but allows us a further insight into the problem.

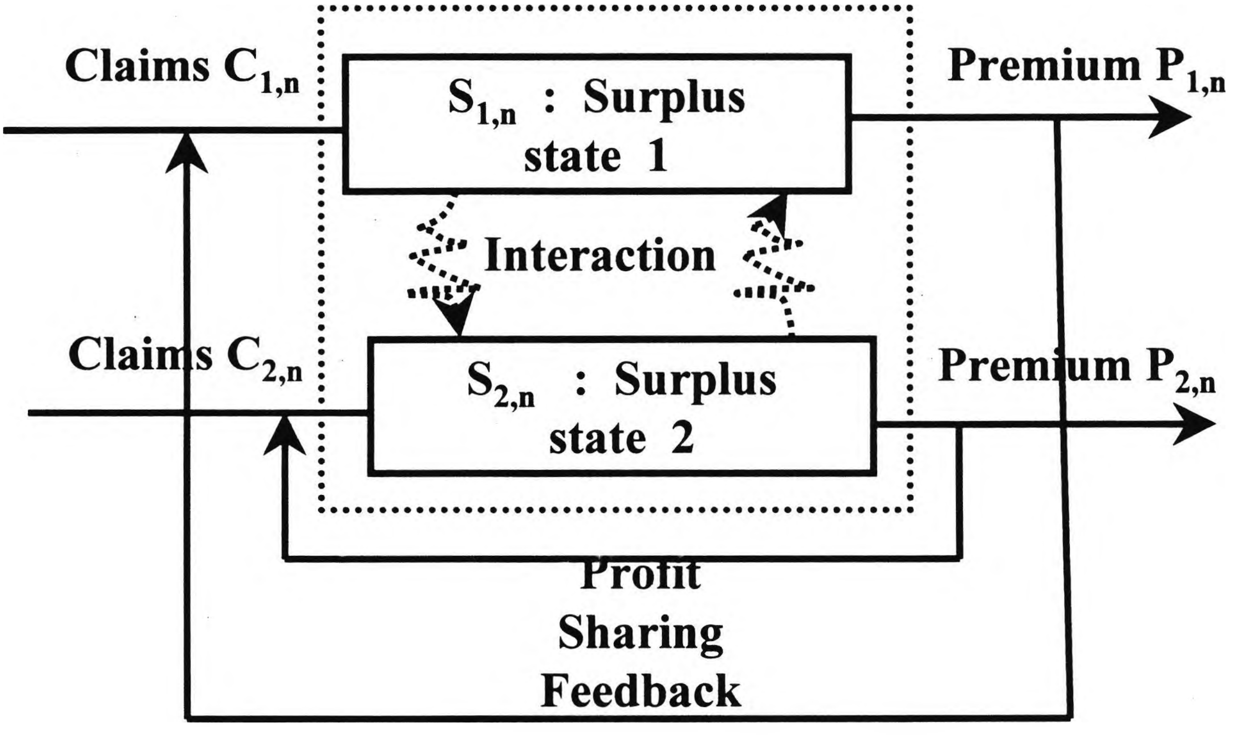


Diagram (5.14.1)

Looking at the diagram above we may identify two directions of smoothing actions. The parallel one which is the profit sharing feedback operating for each of the two subsystems and the vertical one which is determined by the interaction procedure between the two surpluses of the companies.

Furthermore and in order to facilitate the calculations in the next sections we shall produce another model, Model IV assuming that both companies are identical with respect to operational parameters i.e.

$$e_1 = e_2 = e \quad , \quad R_1 = R_2 = R \quad , \quad F_1 = F_2 = F \quad , \quad \varepsilon_1 = \varepsilon_2 = \varepsilon$$

As regards the "harmonization matrix" Λ we shall assume a simple pattern where each company passes exactly the same percentage (λ) of its surplus fund to the other company i.e.

$$\Lambda = \begin{bmatrix} 1-\lambda & \lambda \\ \lambda & 1-\lambda \end{bmatrix}$$

The assumptions above may be reasonable as the multinational networks are composed of (more or less) equivalent (with respect to operational matters) companies. The assumption also is necessary in order to obtain analytical solutions and results.

Numerical Examples

Finally we shall present three numerical examples of Model IV investigating the effect of the interaction factor (λ). We use three values of (λ): a small one (5%), a medium one (50%) and a large one (95%) with the same set of other assumptions i.e.

(1) Input signals

Subsystem 1 : zero (0,0,0,...)

Subsystem 2 : spike (0,1,0,...)

(2) Expense factor (e) : 80%

(3) Interest factor (R) : 1.10

(4) Inflation factor (F) : 1.04

consequently $M'=F^2+F^3=2.21$

We obtain tables (5.14.1), (5.14.2), (5.14.3) and diagrams (5.14.2), (5.14.3) and (5.14.4). As we can observe the large value of (λ) produces oscillations while the small value does not produce oscillations.

In section (5.18) we shall prove formally the result above i.e. the larger the value of (λ) the larger the oscillations. By general reasoning, we could argue that the less intervention (i.e. small value of (λ)) into the system the less the confusion (oscillations) caused in it. Or in other words we should let each subsystem to manage its own problems and not transfer (the problem) using large value of (λ) to the other subsystem.

TABLE (5.14.1)

Input Signal	*Spike (0,0,0,0,...)*	*subsystem 1*
Input Signal	*Zero (0,1,0,0,...)*	*subsystem 2*
Expense (e)	*80%*	
Interest (R)	*1,10*	
Infflation (F)	*1,04*	
(M/)	*2,21*	
Interaction (λ)	*0,05*	

	subsystem 1	*subsystem 2*	*subsystem 1*	*subsystem 2*	*Total System*
Feedback (ε)			*20%*	*20%*	
Time	*Claims*	*Claims*	*Accumulated surplus*	*Accumulated surplus*	*Accumulated surplus*
0	0	0	0,0000	0,0000	0,0000
1	0	1	0,0000	-1,0000	-1,0000
2	0	0	-0,0550	-1,0450	-1,1000
3	0	0	-0,1150	-0,4449	-0,5598
4	0	0	-0,1358	0,2058	0,0700
5	0	0	-0,1122	0,2788	0,1666
6	0	0	-0,0802	0,2522	0,1720
7	0	0	-0,0520	0,2146	0,1626
8	0	0	-0,0297	0,1810	0,1513
9	0	0	-0,0127	0,1532	0,1404
10	0	0	-0,0001	0,1304	0,1303
11	0	0	0,0091	0,1118	0,1208
12	0	0	0,0156	0,0964	0,1121
13	0	0	0,0202	0,0837	0,1039
14	0	0	0,0232	0,0732	0,0964
15	0	0	0,0251	0,0644	0,0894
16	0	0	0,0260	0,0569	0,0829
17	0	0	0,0263	0,0506	0,0769
18	0	0	0,0261	0,0452	0,0713
19	0	0	0,0256	0,0406	0,0662
20	0	0	0,0248	0,0366	0,0614
21	0	0	0,0238	0,0331	0,0569
22	0	0	0,0227	0,0301	0,0528
23	0	0	0,0216	0,0274	0,0490
24	0	0	0,0204	0,0250	0,0454
25	0	0	0,0193	0,0228	0,0421
26	0	0	0,0181	0,0209	0,0391
27	0	0	0,0170	0,0192	0,0362
28	0	0	0,0159	0,0177	0,0336
29	0	0	0,0149	0,0163	0,0312
30	0	0	0,0139	0,0150	0,0289
31	0	0	0,0130	0,0138	0,0268
32	0	0	0,0121	0,0128	0,0249
33	0	0	0,0113	0,0118	0,0231
34	0	0	0,0105	0,0109	0,0214
35	0	0	0,0098	0.0101	0,0198
36	0	0	0,0091	0,0093	0,0184
37	0	0	0,0084	0,0086	0,0171
38	0	0	0,0078	0,0080	0,0158
39	0	0	0,0073	0,0074	0,0147
40	0	0	0,0068	0,0069	0,0136

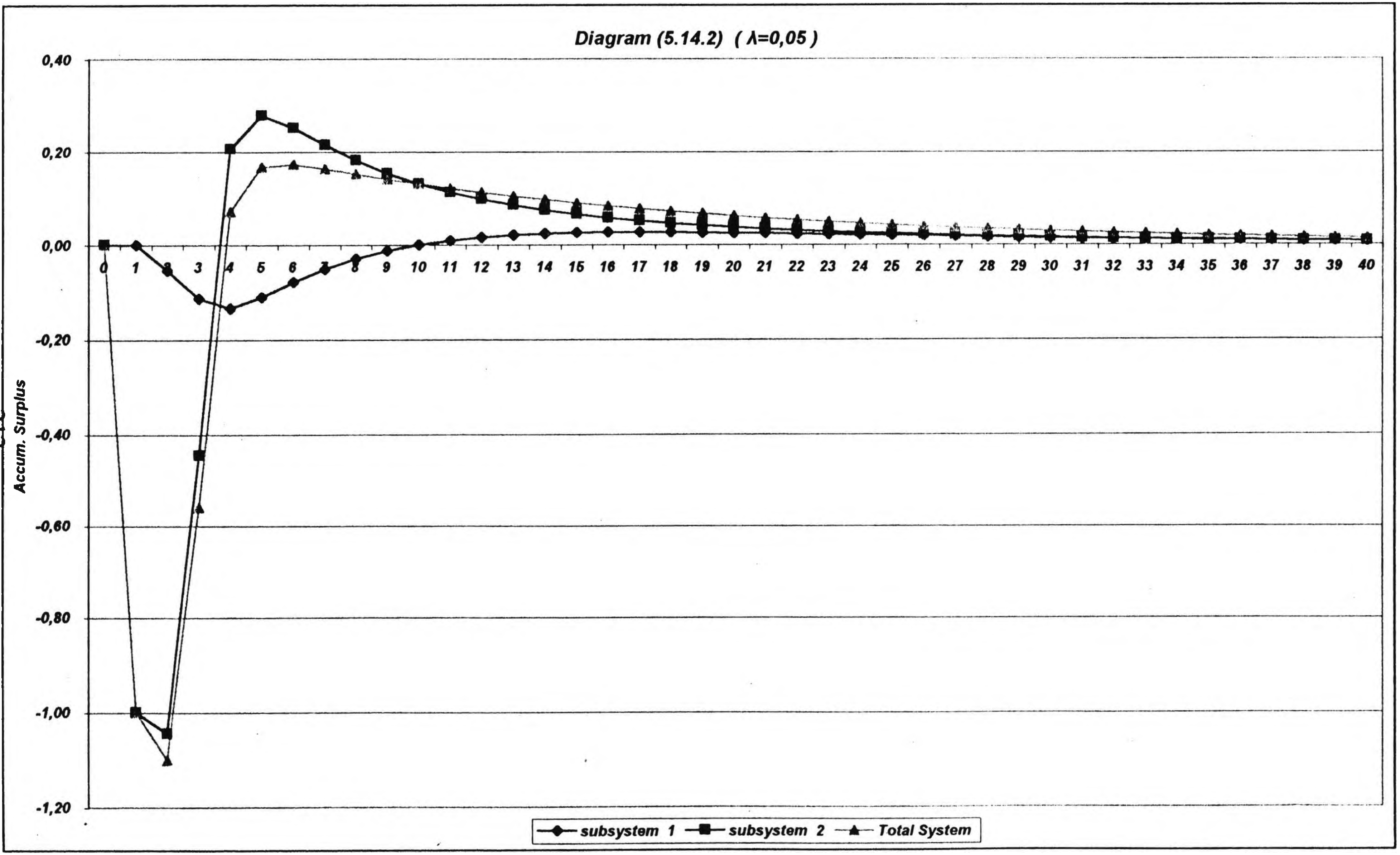


Diagram (5.14.2) (λ=0,05)
Accum. Surplus
0,40
0,20
0,00
-0,20
-0,40
-0,60
-0,80
-1,00
-1,20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
subsystem 1
subsystem 2
Total System

TABLE (5.14.2)

Input Signal	*Spike (0,0,0,0,...)*	*subsystem 1*
Input Signal	*Zero (0,1,0,0,...)*	*subsystem 2*
Expense (e)	*80%*	
Interest (R)	*1,10*	
Infflation (F)	*1,04*	
(M/)	*2,21*	
Interaction (λ)	*0,5*	

	subsystem 1	*subsystem 2*	*subsystem 1*	*subsystem 2*	*Total System*
Feedback (ε)			*20%*	*20%*	
Time	*Claims*	*Claims*	*Accumulated surplus*	*Accumulated surplus*	*Accumulated surplus*
0	0	0	0,0000	0,0000	0,0000
1	0	1	0,0000	-1,0000	-1,0000
2	0	0	-0,5500	-0,5500	-1,1000
3	0	0	-0,6050	0,0452	-0,5598
4	0	0	-0,2199	0,2899	0,0700
5	0	0	0,1353	0,0313	0,1666
6	0	0	0,1268	0,0452	0,1720
7	0	0	0,0730	0,0896	0,1626
8	0	0	0,0691	0,0822	0,1513
9	0	0	0,0716	0,0689	0,1404
10	0	0	0,0662	0,0641	0,1303
11	0	0	0,0602	0,0606	0,1208
12	0	0	0,0559	0,0562	0,1121
13	0	0	0,0520	0,0519	0,1039
14	0	0	0,0482	0,0482	0,0964
15	0	0	0,0447	0,0447	0,0894
16	0	0	0,0415	0,0415	0,0829
17	0	0	0,0385	0,0385	0,0769
18	0	0	0,0357	0,0357	0,0713
19	0	0	0,0331	0,0331	0,0662
20	0	0	0,0307	0,0307	0,0614
21	0	0	0,0285	0,0285	0,0569
22	0	0	0,0264	0,0264	0,0528
23	0	0	0,0245	0,0245	0,0490
24	0	0	0,0227	0,0227	0,0454
25	0	0	0,0211	0,0211	0,0421
26	0	0	0,0195	0,0195	0,0391
27	0	0	0,0181	0,0181	0,0362
28	0	0	0,0168	0,0168	0,0336
29	0	0	0,0156	0,0156	0,0312
30	0	0	0,0145	0,0145	0,0289
31	0	0	0,0134	0,0134	0,0268
32	0	0	0,0124	0,0124	0,0249
33	0	0	0,0115	0,0115	0,0231
34	0	0	0,0107	0,0107	0,0214
35	0	0	0,0099	0,0099	0,0198
36	0	0	0,0092	0,0092	0,0184
37	0	0	0,0085	0,0085	0,0171
38	0	0	0,0079	0,0079	0,0158
39	0	0	0,0073	0,0073	0,0147
40	0	0	0,0068	0,0068	0,0136

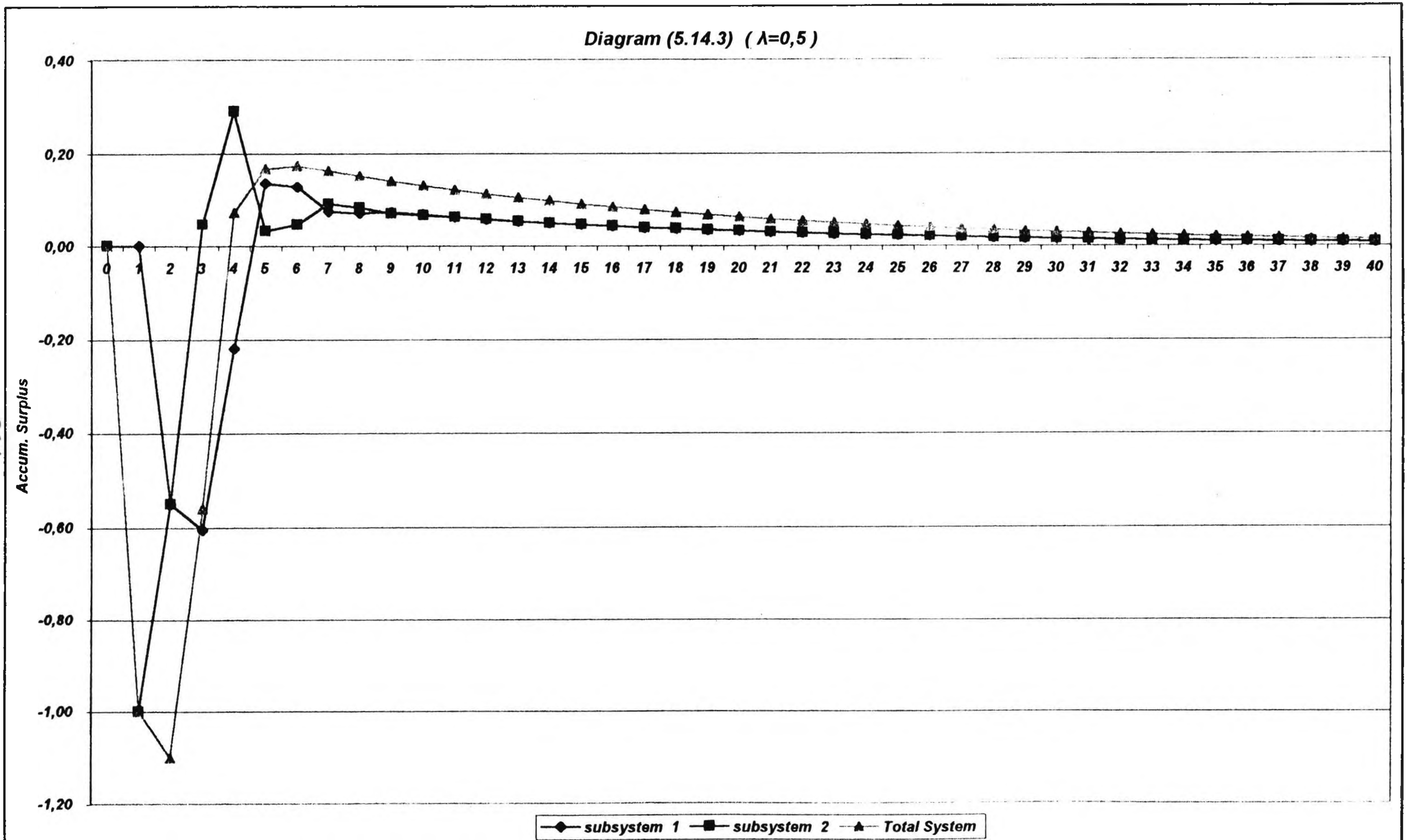


Diagram (5.14.3) (λ=0,5)
0,40
0,20
0,00
-0,20
-0,40
-0,60
-0,80
-1,00
-1,20
Accum. Surplus
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
subsystem 1
subsystem 2
Total System

TABLE (5.14.3)

Input Signal	*Spike (0,0,0,0,...)*	*subsystem 1*
Input Signal	*Zero (0,1,0,0,...)*	*subsystem 2*
Expense (e)	*80%*	
Interest (R)	*1,10*	
Infflation (F)	*1,04*	
(M/)	*2,21*	
Interaction (λ)	*0,95*	

	subsystem 1	*subsystem 2*	*subsystem 1*	*subsystem 2*	*Total System*
Feedback (ε)			*20%*	*20%*	
			Accumulated	*Accumulated*	*Accumulated*
Time	*Claims*	*Claims*	*surplus*	*surplus*	*surplus*
0	0	0	0,0000	0,0000	0,0000
1	0	1	0,0000	-1,0000	-1,0000
2	0	0	-1,0450	-0,0550	-1,1000
3	0	0	-0,1150	-0,4449	-0,5598
4	0	0	-0,3040	0,3740	0,0700
5	0	0	0,3925	-0,2259	0,1666
6	0	0	-0,1659	0,3379	0,1720
7	0	0	0,2812	-0,1186	0,1626
8	0	0	-0,0819	0,2333	0,1513
9	0	0	0,1943	-0,0538	0,1404
10	0	0	-0,0324	0,1627	0,1303
11	0	0	0,1372	-0,0163	0,1208
12	0	0	-0,0043	0,1164	0,1121
13	0	0	0,0995	0,0045	0,1039
14	0	0	0,0108	0,0856	0,0964
15	0	0	0,0741	0,0153	0,0894
16	0	0	0,0184	0,0646	0,0829
17	0	0	0,0566	0,0203	0,0769
18	0	0	0,0214	0,0500	0,0713
19	0	0	0,0443	0,0218	0,0662
20	0	0	0,0218	0,0395	0,0614
21	0	0	0,0354	0,0215	0,0569
22	0	0	0,0209	0,0319	0,0528
23	0	0	0,0288	0,0202	0,0490
24	0	0	0,0193	0,0261	0,0454
25	0	0	0,0237	0,0184	0,0421
26	0	0	0,0174	0,0216	0,0391
27	0	0	0,0198	0,0165	0,0362
28	0	0	0,0155	0,0181	0,0336
29	0	0	0,0166	0,0146	0,0312
30	0	0	0,0137	0,0153	0,0289
31	0	0	0,0140	0,0128	0,0268
32	0	0	0,0119	0,0129	0,0249
33	0	0	0,0119	0,0111	0,0231
34	0	0	0,0104	0,0110	0,0214
35	0	0	0,0102	0,0097	0,0198
36	0	0	0,0090	0,0094	0,0184
37	0	0	0,0087	0,0084	0,0171
38	0	0	0,0078	0,0080	0,0158
39	0	0	0,0074	0,0073	0,0147
40	0	0	0,0067	0,0069	0,0136

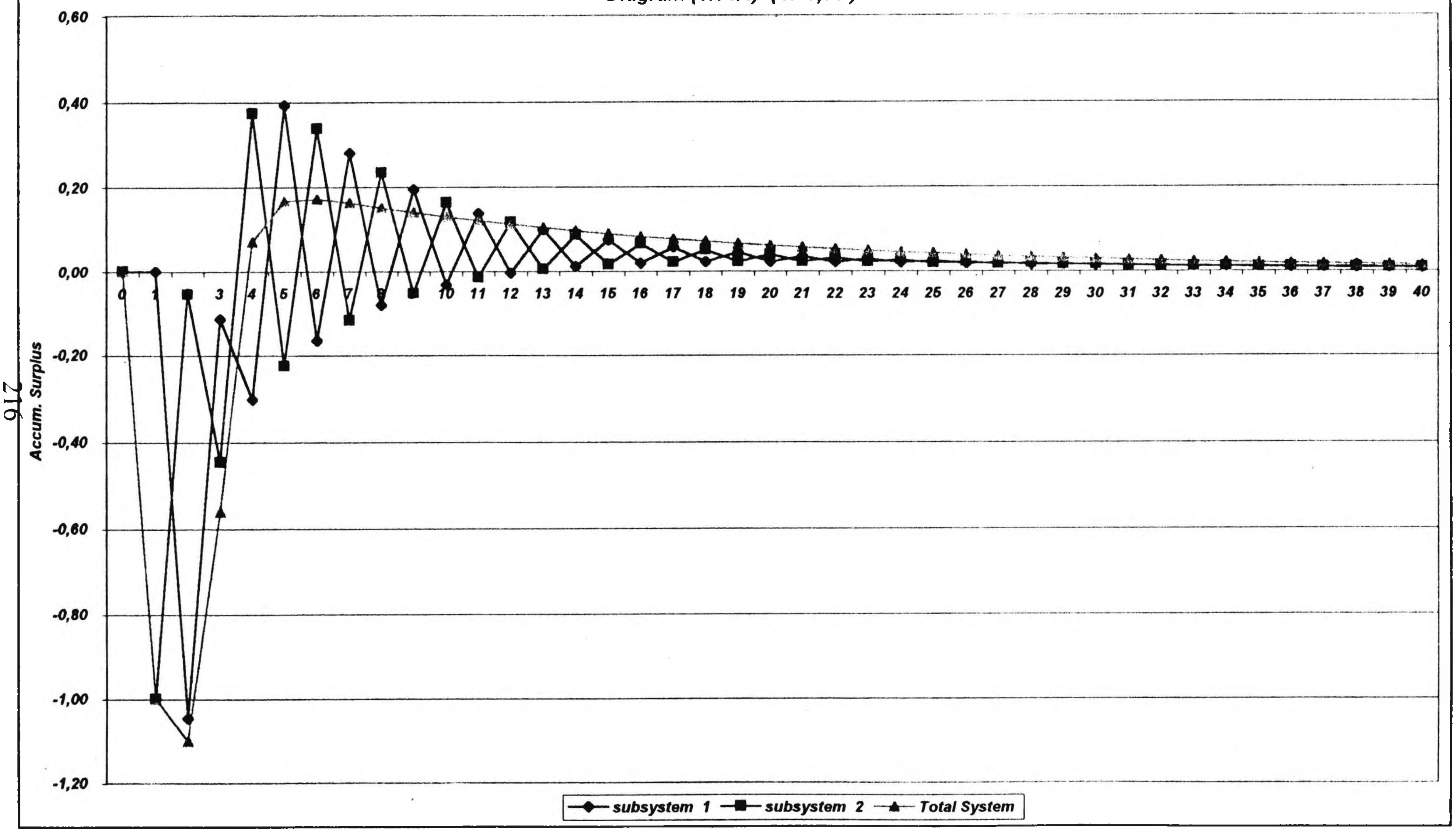


Diagram (5.14.4) (λ=0,95)

5.15 General Solution of the system and output responses to certain input signals (Model IV)

Having obtained the standard format of the dynamic system we may use the standard solution method described in section (2.8.). Hence,

$$\left.\begin{aligned} \underline{x}_n &= A^n \underline{x}_0 + \sum_{k=0}^{n-1} A^k B \underline{u}_{n-k-1}, \quad n=1,2,\ldots \\ \underline{y}_n &= CA^{n-1}\underline{x}_0 + CA^{n-2}B\underline{u}_0 + \ldots + CB\underline{u}_{n-2} + D\underline{u}_n, \quad n=1,2,\ldots \end{aligned}\right\} \tag{5.15.1}$$

As we observe the solution depends firstly on matrix A and its power series (A^n, n=0,1,2,...) and secondly on the other matrices B,C,D and the input signal vectors $\underline{u}_0, \underline{u}_1, \underline{u}_2, \ldots, \underline{u}_n, \ldots$

The calculation of A^n, n=0,1,2,... has explicitly been discussed in chapter 2 and appendix II. The basic target is to determine the diagonial (or Jordan) form of A matrix finding the eigenvalues and the respective eigenvectors.

ρ eigenvalue of matrix $A \Leftrightarrow \det(\rho I - A) = 0$ i.e.

$$\begin{vmatrix} \rho - R(1-\lambda) & e\varepsilon & -R\lambda & 0 \\ -1 & \rho & 0 & 0 \\ -R\lambda & 0 & \rho - R(1-\lambda) & e\varepsilon \\ 0 & 0 & -1 & \rho \end{vmatrix} = 0 \Leftrightarrow$$

(developing the determinant across the second row)

$$\Leftrightarrow (-1)(-1)\begin{vmatrix} e\varepsilon & -R\lambda & 0 \\ 0 & \rho - R(1-\lambda) & e\varepsilon \\ 0 & -1 & \rho \end{vmatrix} + (-1)^2 \rho \begin{vmatrix} \rho - R(1-\lambda) & -R\lambda & 0 \\ -R\lambda & \rho - R(1-\lambda) & e\varepsilon \\ 0 & -1 & \rho \end{vmatrix} = 0$$

$$\Leftrightarrow e \cdot \varepsilon \begin{vmatrix} \rho - R(1-\lambda) & e\varepsilon \\ -1 & \rho \end{vmatrix} + \rho[\rho - R(1-\lambda)]\begin{vmatrix} \rho - R(1-\lambda) & e\varepsilon \\ -1 & \rho \end{vmatrix} + \rho R\lambda \begin{vmatrix} -R\lambda & 0 \\ -1 & \rho \end{vmatrix} = 0$$

$$\Leftrightarrow [\rho[\rho - R(1-\lambda)] + e\varepsilon]\cdot[\rho[\rho - R(1-\lambda)] + e\varepsilon] - [\lambda\rho R]^2 = 0$$

$$\Leftrightarrow [\rho^2 - R(1-\lambda)\rho + e\varepsilon - \lambda R\rho]\cdot[\rho^2 - R(1-\lambda)\rho + e\varepsilon + \lambda R\rho] = 0$$

$$\Leftrightarrow [\rho^2 - R\cdot\rho + e\varepsilon]\cdot[\rho^2 - R(1-2\lambda)\cdot\rho + e\varepsilon] = 0$$

$$\Leftrightarrow \rho^2 - R\rho + e\varepsilon = 0 \quad \text{or} \quad \rho^2 - R(1-\varepsilon\lambda)\rho + e\varepsilon = 0$$

$$\Leftrightarrow \rho_{1,2} = \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2}, \quad \rho_{3,4} = \frac{R(1-2\lambda) \pm \sqrt{R^2(1-2\lambda)^2 - 4e\varepsilon}}{2}$$

Hence, there are four eigenvalues $\rho_1,\rho_2,\rho_3,\rho_4$ which depend (a) the first two on R,e and ε parameters and (b) the second two on R,e,ε and λ parameters. Finally (and after obtaining the eigenvectors) we may rewrite the solution of the dynamic system in the scalar form.

$S_{1,n}=g_1(\ldots)$, $S_{2,n}=g_2(\ldots)$, $P_{1,n}=g_3(\ldots)$, $P_{2,n}=g_4(\ldots)$

where g_1,g_2,g_3,g_4 are real functions of the parameters $\rho_1^n,\rho_2^n,\rho_3^n,\rho_4^n$, (n=0,1,2,...) and R,e,$\varepsilon$,$\lambda$,F.

Closing this section we shall examine the behaviour of the system with respect to three types of input signals (spike, step, sine). Assuming also zero initial conditions $\underline{x}_0 = \underline{0}$ and $\underline{y}_0 = 0$ for any situation.

Spike

We shall assume that a spike signal appears as the input of the first subsystem while the second subsystem has a zero input i.e.

$$C_{1,n} = \begin{cases} 0, & n \neq 0 \\ 1, & n = 0 \end{cases} \qquad C_{2,n} = 0 \qquad \forall n \in \mathbb{Z}$$

then we may obtain the input vector $\underline{u}_n$ as below

$$\underline{u}_0 = \begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}, \ \underline{u}_1 = \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\end{bmatrix}, \ \underline{u}_2 = \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\end{bmatrix}, \ \underline{u}_3 = \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix}, \ \underline{u}_n = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \quad n \geq 4$$

Substituting $\underline{u}_n$ n=0,1,2,... in the general solution we obtain

$$\underline{x}_n = A^n \underline{x}_0 + A^{n-1}B\underline{u}_0 + A^{n-2}B\underline{u}_1 + A^{n-3}B\underline{u}_2 + A^{n-4}B\underline{u}_3$$

$$\underline{y}_n = CA^{n-1}\underline{x}_0 + CA^{n-2}B\underline{u}_0 + CA^{n-3}B\underline{u}_1 + CA^{n-4}B\underline{u}_2 + CA^{n-5}B\underline{u}_3 \quad (n \geq 5)$$

We may continue our calculations so,

$$B\underline{u}_0 = \begin{bmatrix}-1\\0\\0\\0\end{bmatrix}, B\underline{u}_1 = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}, B\underline{u}_2 = \begin{bmatrix}\frac{F^2}{M'}\\0\\0\\0\end{bmatrix}, B\underline{u}_3 = \begin{bmatrix}\frac{F^3}{M'}\\0\\0\\0\end{bmatrix}$$

and since $\underline{x}_0 = \underline{0}$ the solution takes the form

$$\underline{x}_n = A^{n-1} \cdot \begin{bmatrix}-1\\0\\0\\0\end{bmatrix} + A^{n-3} \cdot \begin{bmatrix}\frac{F^2}{M'}\\0\\0\\0\end{bmatrix} + A^{n-4} \cdot \begin{bmatrix}\frac{F^3}{M'}\\0\\0\\0\end{bmatrix}$$

$$\underline{y}_n = CA^{n-2} \cdot \begin{bmatrix}-1\\0\\0\\0\end{bmatrix} + CA^{n-4} \cdot \begin{bmatrix}\frac{F^2}{M'}\\0\\0\\0\end{bmatrix} + CA^{n-5} \cdot \begin{bmatrix}\frac{F^3}{M'}\\0\\0\\0\end{bmatrix}$$

Finally we only need to calculate

A^k, k=n-1, n-2, n-3, n-4, n-5 and substitute them into the formulae above.

Obviously, from the final format of the state $\underline{x}_n$ and output $\underline{y}_n$ the system will asymptotically converge to zero (as n increases) assuming that all the magnitudes of the eigenvalues of A are less than the unity. This is true because only a finite number of powers of A is involved in the formulae (i.e. A^{n-1}, A^{n-3}, A^{n-4} for $\underline{x}_n$ and A^{n-2}, A^{n-4}, A^{n-5} for $\underline{y}_n$).

Step

We assume a step signal for the first input variable i.e.

$$C_{1,n} = \begin{cases} 0, & n < 0 \\ 1, & n \geq 0 \end{cases} \qquad C_{2,n} = 0 \qquad \forall n \in \mathbb{Z} \quad \text{then,}$$

$$\underline{u}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \underline{u}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \underline{u}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \underline{u}_n = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad n \geq 3,$$

and consequently

$$B\underline{u}_0 = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B\underline{u}_1 = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B\underline{u}_2 = \begin{bmatrix} \frac{F^2}{M'} - 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, B\underline{u}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad n \geq 3$$

(The last equality holds because $M' = F^2 + F^3$, so $B\underline{u}_n = \underline{0} \quad n \geq 3$).

Now the general solution may be written in the form

$$\underline{x}_n = \left[A^{n-1} + A^{n-2}\right] \cdot \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + A^{n-3} \cdot \begin{bmatrix} \frac{F^2}{M'} - 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\underline{y}_n = C \cdot \left[A^{n-2} + A^{n-3}\right] \cdot \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + CA^{n-4} \cdot \begin{bmatrix} \frac{F^2}{M'} - 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad , \quad n \geq 5$$

Again as for the spike input we observe that the format of $\underline{x}_n$ and $\underline{y}_n$ contains a finite number of powers of Matrix A. So assuming again eigenvalues with magnitude less than the unity we may obtain asymptotic stability and convergency to initial conditions of the system (zero state and output).

Sine

An interesting assumption about the input variable is the sine signal. This model of input signal may be the more realistic in some cases as the sine waves represent the underwriting cycles which appear in the insurance market of each country. So we define

$$C_{1,n} = \sin(\omega_1 n + \varphi_1) \ , \ C_{2,n} = \sin(\omega_2 n + \varphi_2) \quad n \geq 0$$

$$\omega_1 = \omega_2 = \pi \quad , \quad \varphi_1 = -\frac{\pi}{6} \quad , \quad \varphi_2 = \frac{\pi}{2}$$

n	0	1	2	3	4
$C_{1,n}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	...
$C_{2,n}$	1	-1	1	-1	...

Consequently the input vectors,

$$\underline{u}_0 = \begin{bmatrix} -0.5 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \underline{u}_1 = \begin{bmatrix} 0.5 \\ -0.5 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ \underline{u}_2 = \begin{bmatrix} -0.5 \\ 0.5 \\ -05 \\ 0 \\ 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \ \underline{u}_k = \begin{bmatrix} 0,5 \\ -0,5 \\ 0,5 \\ -0,5 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}, \ \underline{u}_\mu = \begin{bmatrix} -0,5 \\ 0,5 \\ -0,5 \\ 0,5 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

where k=2n+1, μ=2(n+1), n≥1

$$B\underline{u}_0 = \begin{bmatrix} 0,5 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \ B\underline{u}_1 = \begin{bmatrix} -0,5 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \ B\underline{u}_2 = \begin{bmatrix} 0,5\left(1-\dfrac{F^2}{M'}\right) \\ 0 \\ -\left(1-\dfrac{F^2}{M'}\right) \\ 0 \end{bmatrix}$$

$$B\underline{u}_k = \begin{bmatrix} -0,5\left(1-\dfrac{F^2}{M'}+\dfrac{F^3}{M'}\right) \\ 0 \\ 1-\dfrac{F^2}{M'}+\dfrac{F^3}{M'} \\ 0 \end{bmatrix}, \ B\underline{u}_\mu = \begin{bmatrix} 0,5\left(1-\dfrac{F^2}{M'}+\dfrac{F^3}{M'}\right) \\ 0 \\ -\left[1-\dfrac{F^2}{M'}+\dfrac{F^3}{M'}\right] \\ 0 \end{bmatrix}$$

where k=2n+1, μ=2(n+1) n≥1.

As we observe for the vectors calculated before

$$B\underline{u}_1 = -B\underline{u}_0 \quad \text{and} \quad B\underline{u}_{k+1} = -B\underline{u}_k \ , \quad k=2n+1,\ n\geq 1$$

Now assuming again zero initial condition i.e. $\underline{x}_0 = \underline{0}$ and substituting equations mentioned above we obtain the general solution for the state of the system

$$\underline{x}_n = A^{n-1}B\underline{u}_0 - A^{n-2}B\underline{u}_0 + A^{n-3}B\underline{u}_2 + A^{n-4}B\underline{u}_3 - A^{n-5}B\underline{u}_3 + \ldots + (-1)^{n-1}AB\underline{u}_3 + (-1)^n B\underline{u}_3$$

Rearranging the terms of the equation above we obtain

$$\underline{x}_n = A^{n-2}(A-I)B\underline{u}_0 + A^{n-3}B\underline{u}_2 + \left(A^{n-4} - A^{n-3} + \ldots + (-1)^n I\right)B\underline{u}_3$$

Now we shall distinguish the values for odd and even numbers i.e.

n = odd number

$$\underline{x}_n = A^{n-2}(A-I)B\underline{u}_0 + A^{n-3}B\underline{u}_2 + \left(I + A^2 + \ldots + A^{n-3}\right)(A-I)B\underline{u}_3$$

n = even number

$$\underline{x}_n = A^{n-2}(A-I)\,B\underline{u}_0 + A^{n-3}B\underline{u}_2 + \left[(A-I)\left(A + A^3 + \ldots + A^{n-3}\right) + I\right]B\underline{u}_3$$

If we want to examine the behaviour of the system for large values of n (as n goes to infinity) we may prove that the sequence of $\|\underline{x}_n\|$ converges to a certain limit and consequently the sequence of $\underline{x}_n$ converges respectively.

If we recall the norms defined in section (4.21) and define

$$\xi = |||A||| \ , \ \xi' = |||B|||$$

we obtain

n = odd number

$$\|\underline{x}_n\| \le \xi^{n-2}(\xi+1)\xi'\|\underline{u}_0\| + \xi^{n-3}\xi'\|\underline{u}_2\| + (1+\xi^2+\ldots+\xi^{n-3})(\xi+1)\xi'\|\underline{u}_3\| =$$

$$\xi^{n-2}(\xi+1)\xi'\|\underline{u}_0\| + \xi^{n-3}\xi'\|\underline{u}_2\| + \frac{(\xi^2)^{\frac{n-3}{2}+1}-1}{\xi^2-1}(\xi+1)\xi'\|\underline{u}_3\| =$$

$$\xi^{n-2}(\xi+1)\xi'\|\underline{u}_0\| + \xi^{n-3}\xi'\|\underline{u}_2\| + \frac{(\xi^2)^{\frac{n-3}{2}}-1}{\xi-1}\xi'\|\underline{u}_3\|$$

If we choose the parameters of the system such that $\xi = |||A||| < 1$ then the sequence of $\|\underline{x}_n\|$ is bounded from another sequence (see the last inequality) which converges and so the sequence of $\|\underline{x}_n\|$ (for odd numbers) is bounded

n = even number

Similarly as for the case where n is an odd number we may prove that

$$\|\underline{x}_n\| \le \xi^{n-2}(\xi+1)\xi'\|\underline{u}_0\| + \xi^{n-3}\xi'\|\underline{u}_2\| + [(\xi+1)(\xi+\xi^2+\ldots+\xi^{n-3})+1]\xi'\|\underline{u}_3\|$$

and consequently (see the argument for an odd number) the sequence of $\|x_n\|$ for even number is bounded.

Finally we obtain that sequence $\|x_n\|$ converges and consequently the sequence x_n also converges to a certain limit as n goes to infinity.

5.16 Stability Analysis of Model IV

One of the main questions for any dynamic system is the existence of some kind of stability. With reference to the analysis of section (2.15) we have the following development.

First of all we should find the equilibrium points $\underline{x}_e$. As we know, $\underline{x}_e$ equilibrium point $\Leftrightarrow A\underline{x}_e = \underline{x}_e$.

But $$\det(A)=\begin{vmatrix} R(1-\lambda) & -e\varepsilon & R\lambda & 0 \\ 1 & 0 & 0 & 0 \\ R\lambda & 0 & R(1-\lambda) & -e\varepsilon \\ 0 & 0 & 1 & 0 \end{vmatrix} = -\begin{vmatrix} -e\varepsilon & R\lambda & 0 \\ 0 & R(1-\lambda) & -e\varepsilon \\ 0 & 1 & 0 \end{vmatrix} = (e\varepsilon)^2$$

Given that $e\neq 0$ and $\varepsilon\neq 0$ then $\det(A)\neq 0$ and consequently A is non singular and the system $A\underline{x}_e = \underline{x}_e$ has a unique solution, i.e. the trivial $\underline{x}_e = \underline{0}$ (since $A\cdot\underline{0}=\underline{0}$) .

Now we shall examine whether the zero equilibrium point is also a Liapunov or asymptotic stable point. As we have seen in section (2.15) and the relevant stability criterion, we should examine the magnitude of the eigenvalues of A against the magnitude of the unity (less or equal to the unity).

The eigenvalues of A matrix have been found in section (5.15) as follows

$$\rho_{1,2} = \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2} \quad \text{and} \quad \rho_{3,4} = \frac{R(1-2\lambda) \pm \sqrt{R^2(1-2\lambda)^2 - 4e\varepsilon}}{2}$$

We shall gain further insight for the magnitude of the roots if we plot them in the xy-plane and identify the area (of root loci) which lies within the unit circle (see Diagram 5.16.1).

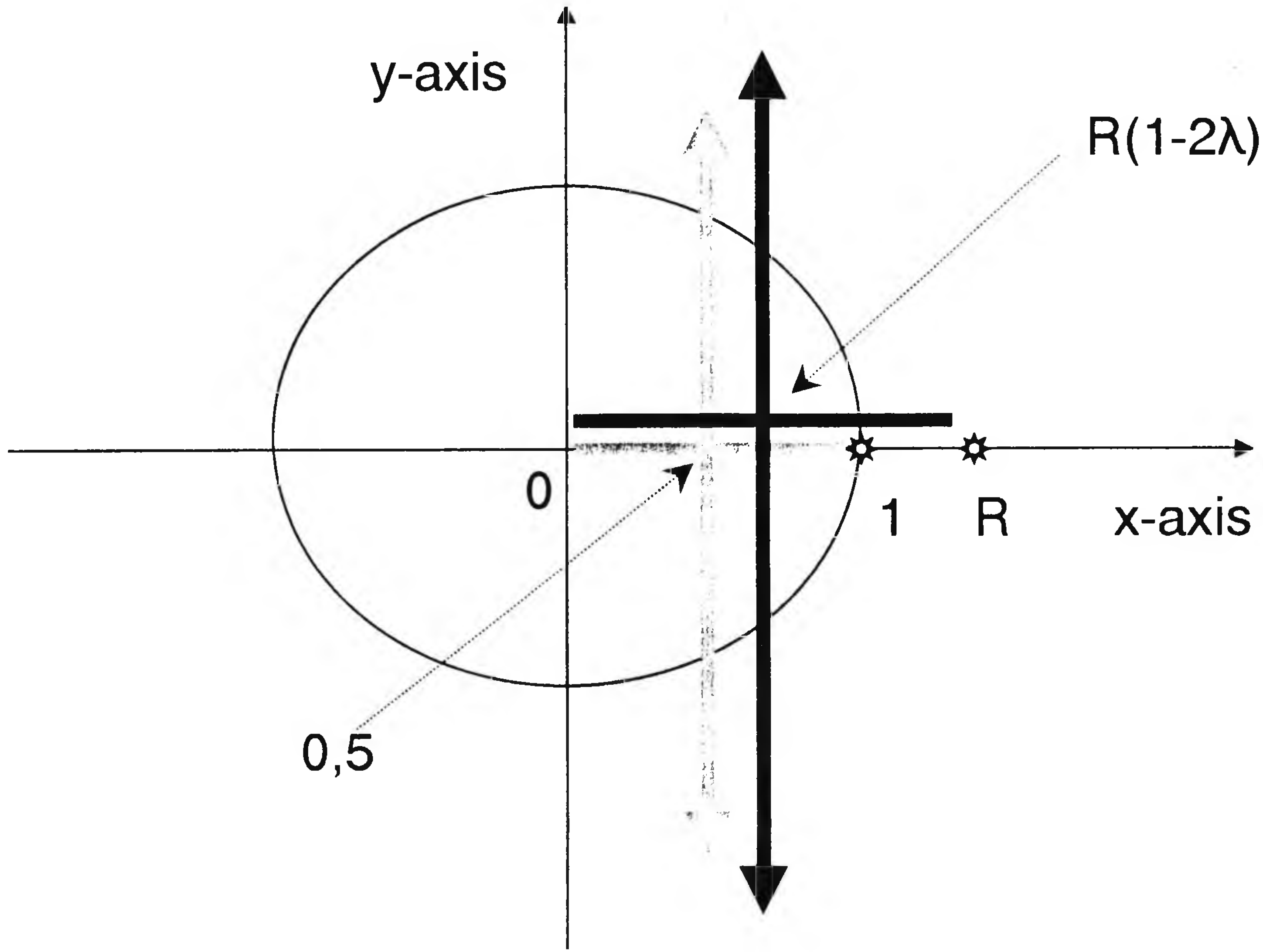


Diagram (5.16.1)

All the roots may be plotted on two crosses for constant R,λ and as the product $e\varepsilon$ varies from zero to infinity.

The characteristics of these crosses are the following:

(a) The centre are the breakaway points with x-axis lie on $\frac{R}{2}$ and $\frac{R(1-2\lambda)}{2}$ for ρ_1,ρ_2 and ρ_3,ρ_4 respectively.

(b) The first axis of each cross lies on the x-axis from 0 up to R for ρ_1,ρ_2 roots and from 0 up to R(1-2λ) for ρ_3,ρ_4 roots.

(c) The second axes are parallel to y-axis (continuing up to infinity as $e\varepsilon$ goes to infinity).

Interpreting also the diagram with respect to the stability of the system we may comment that: For a given pair of R,λ and constant e then the extreme choices for ε (i.e. zero or infinity, infinity means too large) will produce an unstable system.

We shall now find conditions for the breakpoints of the unit circle, but before we shall determine the intervals for the typical values of the parameters involved

$\varepsilon \in [0,1]$ $\qquad$ $R \in [1\ ,\ 1.10], \lambda \in [0.05\ ,\ 1], e \in [0.5\ ,\ 1]$

Stability Conditions

• For the first two roots ρ_1 and ρ_2 we shall distinguish two situations (real and complex roots)

Real roots ρ_1,ρ_2 i.e. R^2-4eε>0

$$|\rho_{1,2}| < 1 \Leftrightarrow \left|\frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2}\right| \leq 1 \Leftrightarrow$$

(the root with the negative sign normally is less than 1)

$$\Leftrightarrow \frac{R+\sqrt{R^2-4e\varepsilon}}{2} \leq 1 \Leftrightarrow \sqrt{R^2-4e\varepsilon} \leq 2-R \Leftrightarrow$$

$$R^2-4e\varepsilon \leq (2-R)^2 \Leftrightarrow R^2-4e\varepsilon \leq R^2-4R+4$$

Hence, $$e\varepsilon \geq R-1 \qquad (5.16.2)$$

Complex roots ρ_1,ρ_2 i.e. R^2-4eε<0

If R^2-4eε<0 then we have two complex conjugates roots $\rho_{1,2} = \frac{R}{2} \pm i \cdot \frac{\sqrt{4e\varepsilon - R^2}}{2}$

and consequently

$$|\rho_{1,2}| \leq 1 \Leftrightarrow \left[\left(\frac{R}{2}\right)^2 + \left(\frac{\sqrt{4e\varepsilon - R^2}}{2}\right)^2\right]^{\frac{1}{2}} \leq 1 \Leftrightarrow$$

$$\Leftrightarrow \frac{R^2}{4} + \frac{4e\varepsilon - R^2}{4} \leq 1 \Leftrightarrow e\varepsilon \leq 1$$

• For the second two roots ρ_3 and ρ_4 we shall follow the same procedure as above and it may be easily proved that in order to obtain roots with absolute value less than unity

$$R(1-2\lambda)-1<e\varepsilon<1 \qquad (5.16.3)$$

Finally summing up all the work for the magnitute of the roots we obtain the follow. The zero point is a Liapunov stable point if and only if

$$R-1 \leq e\varepsilon \leq 1 \qquad (5.16.4)$$

(since $R(1-2\lambda)-1 < R-1$)

Consequently the zero point is an asymptotic stable point if and only if

$$R-1 < e\varepsilon < 1 \qquad (5.16.5)$$

(The conditions (5.16.4) is similar with (4.17.6)). The only difference arises due to the fact that now there is not investment income for premiums in the year received. The parameter (λ) is not involved in the inequality above but we can see from another point of view how it affects the stability of the system.

The choice of λ determines the centre of the second cross. Large values of λ (near to 100%) move the centre of the cross near to zero and consequently a bigger area of the cross lies in the unit circle, so the probability of a stable system is increased.

5.17 Controllability and Observability properties of Models III, IV

As regards the investigation of these properties we should refer to section (5.7). As we have stated Model III is a restriction of Model I. Consequently, since the Model I has the two properties above (as shown in (5.7)) its obvious that also Model III has these properties too. The same applies for Model IV.

5.18 Optimal design for the parameters of Model IV with respect to speed response and oscillatory form of the solution

Another interesting issue of our problem is the determination of the optimal parameter values according to some specific requirements and criteria. In this section, we shall examine four cases with respect to the oscillatory form of the solution and the required response speed to different input signals.

1st Case: No oscillations

As a first case, we require a solution of the system with no oscillations. The reason why, may be easily found. Oscillations mean that the state and consequently the output will fluctuate round the x-axis. So the premium charged to the policy holder will sometimes increases and other times decreases with a sine wave form. Of course this is not desirable especially when the fluctuation (up and down) is very big.

From the mathematical point of view no oscillations means that all the eigenvalues of A are real numbers. Looking the form of $\rho_1,\rho_2,\rho_3,\rho_4$ we demand

$$\left.\begin{aligned} R^2-4e\varepsilon \geq 0 \\ R^2(1-2\lambda)^2-4e\varepsilon \geq 0 \end{aligned}\right\}$$

and since λ normally lies between zero and unity the conditions above are restricted to one i.e.

$$R^2(1-2\lambda)^2-4e\varepsilon \geq 0$$

2^{nd} Case: Fastest Response

The speed of the state or output response to the different input signals (the minimum time which the system needs to return to zero, equilibrium point) depends on the maximum magnitude of the eigenvalues of A. As we have seen the solution depends on the power series of A (A^n n=0,1,2,...) which depends on the powers of the eigenvalues.

Obviously from the form of the eigenvalues we derive that the minimum values are obtained when the quadratic polynomials have double roots i.e.

$$\left.\begin{aligned} R^2-4e\varepsilon = 0 \\ R^2(1-2\lambda)^2-4e\varepsilon = 0 \end{aligned}\right\}$$

The conditions above may be simultaneously fulfilled if and only if λ=0 or 1, where we obtain the trivial cases with no or full interaction between the two systems and we have two equal double roots (or a root with fourth multiplicity)

$$\rho_{1,2} = \rho_{3,4} = \frac{R}{2}$$

In the real case where $\lambda>0$ then we can not choose parameters making true both equations but we still have to minimize the maximum magnitude of the roots. So we have to choose which equation is most important (producing the maximum magnitude).

We may answer the question above if we observe Diagram (5.16.1). The ρ_1,ρ_2 lies on the right cross while ρ_3,ρ_4 lies on the left cross. We can see that if we demand R^2-4eε=0 then ρ_1,ρ_2 coincides on the centre of the right cross while ρ_3,ρ_4 are complex numbers and lie on the vertical axis of the left cross. So the root with the maximum absolute value is the double root $\rho_1,\rho_2 = \frac{R}{2}$.

The other option of demanding $R^2(1-2\lambda)^2 - 4e\varepsilon = 0$ will produce four eigenvalues as follows. A double root $\rho_3,\rho_4 = \frac{R(1-2\lambda)}{2}$ and two different $\rho_1,\rho_2 = \frac{R \pm \sqrt{R^2 - 4e\varepsilon}}{2}$. Obviously the maximum absolute value exists for root $\rho_2 = \frac{R + \sqrt{R^2 - 4e\varepsilon}}{2}$ which is bigger than $\frac{R}{2}$.

Finally, we may derive the following two proposals.

Fastest Response with oscillations $\Leftrightarrow R^2 - 4e\varepsilon = 0$.

Fastest Response with no oscillations $\Leftrightarrow (1-2\lambda)^2 R^2 - 4e\varepsilon = 0$.

3rd Case: Optimal Response & Oscillations

In the first two cases we have identified an opposite effect between the fastest response and system oscillations. Here, we shall try to find a compromise between them reducing the oscillations down to an acceptable level.

Firstly, we require the fastest response, so we choose ε such that R^2-4eε=0.

From the choice above we obtain

- One double root $\rho_{1,2} = \frac{R}{2}$ and

- Two complex conjugate roots

$$\rho_{3,4} = \frac{R}{2} \cdot \left[(1 - 2\lambda) \pm i \cdot \sqrt{4(\lambda - \lambda^2)} \right]$$

The last pair of the complex form may be written

$$\rho_{3,4} = \frac{R}{2} (\cos\theta \pm i \sin\theta)$$

where $$\tan\theta = \frac{\sqrt{\lambda - \lambda^2}}{0.5 - \lambda}$$

It is obvious that the choice of λ does not affect the absolute value of the roots but only affects the value of angle θ.

The two roots ρ_3 and ρ_4 may be plotted on a circle with a radius of $\frac{R}{2}$. According to the choice of λ the roots run over the circle (see next diagram (5.18.1)).

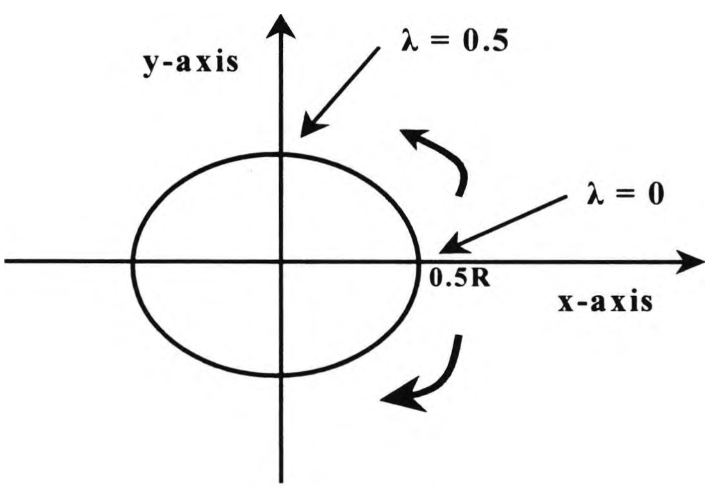


Diagram (5.18.1)

At this point we shall open a small parenthesis and discuss the angle θ in conjunction with the general solution of the system.

If z_1, z_2 are complex conjugates eigenvalues of matrix A of a dynamic system then the general solution will contain a linear combination of the powers of z_1 and z_2 i.e.

$$y_n = \mu_1 z_1^n + \mu_2 z_2^n$$

where $\mu_{1,2} = a(\cos\beta \pm i \sin\beta)$ and $z_{1,2} = z(\cos\theta + i\sin\theta)$

Hence, $y_n = 2az^n \cos(\eta\theta + \beta)$

As we observe from the last equation the angle θ determines the magnitude of oscillations. Consequently the choice of interaction parameter λ will affect the magnitude of oscillations.

If we ignore β as a constant item we can see that the sign of y_n depends on angle θ. So if we require two consecutive values of y_n having the same sign it should be

$$\theta \le \frac{\pi}{4}$$

The last condition may be expressed as

$$\tan\theta \le 1$$

and solving with respect to λ we obtain

$$\tan\theta = \frac{\sqrt{\lambda - \lambda^2}}{0.5 - \lambda} \le 1 \Leftrightarrow \sqrt{\lambda - \lambda^2} \le 0.5 - \lambda \Leftrightarrow$$

$$\overset{(\lambda<0.5)}{\Leftrightarrow} \lambda - \lambda^2 \le (0.5 - \lambda)^2 \Leftrightarrow \lambda - \lambda^2 \le 0.25 - \lambda + \lambda^2 \Leftrightarrow$$

$$2\lambda^2 - 2\lambda + 0.25 \ge 0$$

The roots of quadratic are

$$\lambda_{1,2} = \frac{1 \pm \sqrt{0.5}}{2} \ .$$

The inequality holds outside the roots and since λ<0.5 we obtain

$$\lambda \leq \frac{1-\sqrt{0.5}}{2} \cong 0.1464 \ \ (14.64\%) \ .$$

If we require three consecutive values of the y_n having the same sign it should be

$$\theta \leq \frac{\pi}{6}$$

or equivalently

$$\tan\theta \leq \frac{\sqrt{3}}{3} \ .$$

Solving with respect to λ we obtain

$$\tan\theta = \frac{\sqrt{\lambda-\lambda^2}}{0.5-\lambda} \leq \frac{\sqrt{3}}{3} \Leftrightarrow 3\cdot\sqrt{\lambda-\lambda^2} \leq \sqrt{3}\cdot(0.5-\lambda) \Leftrightarrow$$

$$\overset{(\lambda<0.5)}{\Leftrightarrow} 9\left(\lambda-\lambda^2\right) \leq 3(0.5-\lambda)^2 \Leftrightarrow 3\lambda - 3\lambda^2 \leq 0.25 - \lambda + \lambda^2 \Leftrightarrow$$

$$4\lambda^2 - 4\lambda + 0.25 \geq 0 \ .$$

The roots of quadratic are

$$\lambda_{1,2} = \frac{2 \pm \sqrt{3}}{4} \ .$$

Again the inequality holds outside the roots and since λ<0.5 we obtain

$$\lambda \leq \frac{2-\sqrt{3}}{4} \cong 0.067 \ \ (\cong 6.7\%) \ .$$

Summarizing our results we have the following table

Table (5.18.1)

λ	0	6.7%	14.64%	50%
θ	0	$\pi/3$	$\pi/4$	$\pi/2$

5.19 Optimal reinsurance (Model IV) with respect to the variance of surplus (Comparison with other models of optimal reinsurance, static or dynamic)

Optimal reinsurance is a very large subject and it has been widely discussed in the risk-theoretical literature. The problem is usually tackled using static models i.e. Fix the time period and then divide the total claim amounts into cedant's and reinsurers components in an optimal way (see Daykin et al (1994)).

A dynamic approach has been presented by Rantala (1987) (see section (3.16)) using the concepts of control theory. His target was to minimize the variance of the retained claims while keeping the solvency requirement constant.

As we have seen in the previous sections, our approach does not target the minimum variance of retained claims. Our main effort is to establish a stable system of risk exchange between two insurance partners (not between an insurer and a reinsurer) over a long time period.

We have also attempted to target the fastest response and asymptotic behaviour of the system (return to initial equilibrium state). In the next part of the section we shall try to handle the problem with respect to the variances and get an optimal solution.

Of course our target will be the minimization of the variance of surplus (since there is no claim exchange between the two companies, the incurred claims equals the retained claims).

As we have seen in section (5.5) the surplus vector $\underline{x}_n$ is given from the general equation

$$\underline{x}_n = A^n \underline{x}_0 + \sum_{k=0}^{n-1} A^k B \underline{u}_{n-k-1}, \quad n=1,2,...$$

The first term may disappear if $\underline{x}_0 = \underline{0}$ (zero initial conditions). So, $\underline{x}_n$ is expressed only as a combination of $\underline{u}_k$ k=1,...n-1.

$$\underline{x}_n = B\underline{u}_{n-1} + AB\underline{u}_{n-2} + ... + A^{n-1}B\underline{u}_0, \quad n=1,2,...$$

substituting $H_n=A^n \cdot B$, n=0,1,2,... we obtain

$$\underline{x}_n = H_0 \underline{u}_{n-1} + H_1 \underline{u}_{n-2} + ... + H_{n-1} \underline{u}_0$$

where H_n is a matrix with dimensions 4×8.

Then $$Var(\underline{x}_n) = \begin{bmatrix} Var(s_{1,n}) \\ Var(s_{1,n-1}) \\ Var(s_{2,n}) \\ Var(s_{2,n-1}) \end{bmatrix} = Var\left[H_0 \underline{u}_{n-1} + H_1 \underline{u}_{n-2} + ... + H_{n-1} \underline{u}_0\right]$$

We shall calculate analytically the $Var(\underline{x}_n)$ for n=1,2 and examine its behaviour with respect to λ. For n≥3 we shall provide only the theoretical analysis up to a certain point. We assume $C_{i,n}$ i=1,2 , n=0,±1,±2,... are independent identically random variables with

$$var(C_{i,n}) = \sigma^2 \quad i=1,2 \ , \ n \in \mathbb{Z}$$

and consequently for $\underline{u}_n$ we obtain

$$var(\underline{u}_n) = \sigma^2 \cdot \underline{1} \quad n \in \mathbb{Z} \text{ where } \underline{1} \in \mathbb{R}^{8\times 1}$$

Now for n=1 we obtain

$$\mathrm{var}(\underline{x}_1) = \mathrm{var}(\underline{H}_0 \underline{u}_0) = \mathrm{var}(B\underline{u}_0)$$

$$= \mathrm{var}\left(\begin{bmatrix} -1C_{1,0} + \frac{F^2}{2}C_{1,-2} + \frac{F^3}{2}C_{1,-3} \\ 0 \\ -1C_{2,0} + \frac{F^2}{2}C_{2,-2} + F^3C_{2,-3} \\ 0 \end{bmatrix}\right)$$

taking the variance of each coordinate we have the following result

$$\mathrm{var}(\underline{x}_1) = \sigma^2 \cdot \begin{bmatrix} 1 + \frac{F^4}{M'^2} + \frac{F^6}{M'^2} \\ 0 \\ 1 + \frac{F^4}{M'^2} + \frac{F^6}{M'^2} \\ 0 \end{bmatrix}$$

As we observe the variance of $\underline{x}_1$ does not depend on the choice of λ (depends upon F).

For n=2 we obtain

$$\mathrm{var}(\underline{x}_2) = \mathrm{var}(H_0 \underline{u}_1 + H_1 \underline{u}_0) =$$
$$= \mathrm{var}(B\underline{u}_1 + AB\underline{u}_0)$$

In section (5.7) we have calculated the product A·B. Consequently, we obtain

$$\mathrm{Var}(\underline{x}_2) = \mathrm{Var}\left(\begin{bmatrix} -1C_{1,1} + \frac{F^2}{M'}C_{1,-1} + \frac{F^3}{M'}C_{1,-2} \\ 0 \\ -1C_{2,1} + \frac{F^2}{M'}C_{2,-1} + \frac{F^3}{M'}C_{2,-2} \\ 0 \end{bmatrix} + \right.$$

$$\left. + \begin{bmatrix} (1-\lambda)\left[-RC_{1,1} + \frac{F^2}{M'}RC_{1,-1} + \frac{F^3}{M'}RC_{1,-2}\right] + \lambda\left[-RC_{2,1} + \frac{F^2}{M'}RC_{2,-1} + \frac{F^3}{M'}RC_{2,-2}\right] \\ -1C_{1,1} + \frac{F^2}{M'}C_{1,-1} + \frac{F^3}{M'}C_{2,-2} \\ (1-\lambda)\left[-RC_{1,1} + \frac{F^2}{M'}RC_{1,-1} + \frac{F^3}{M'}RC_{2,-2}\right] + \lambda\left[-RC_{1,1} + \frac{F^2}{M'}RC_{1,-1} + \frac{F^3}{M'}RC_{1,-2}\right] \\ -1C_{2,1} + \frac{F^2}{M'}C_{2,-1} + \frac{F^3}{M'}C_{2,-2} \end{bmatrix}\right)$$

$$\mathrm{Var}(\underline{x}_2)=\begin{bmatrix}\dfrac{\left[(1+(1-\lambda)R)^2+\lambda^2R^2\right]\cdot\left[1+\dfrac{F^4}{M'^2}+\dfrac{F^6}{M'^2}\right]}{1+\dfrac{F^4}{M'^2}+\dfrac{F^6}{M'^2}} \\ \dfrac{\left[(1+(1-\lambda)R)^2+\lambda^2R^2\right]\cdot\left[1+\dfrac{F^4}{M'^2}+\dfrac{F^6}{M'^2}\right]}{1+\dfrac{F^4}{M'^2}+\dfrac{F^6}{M'^2}}\end{bmatrix}$$

As we can see now, the variance of $\underline{x}_2$ depends on λ. If we want the minimum value of $\mathrm{var}(\underline{x}_2)$ with respect to λ we have to minimize the function

$$f(\lambda)=1+(1-\lambda)^2R^2+\lambda^2R^2$$

The minimum if obtained for λ_0 such that

$$\left.\frac{df(\lambda)}{d\lambda}\right|_{\lambda=\lambda_0}=0 \text{ and } \left.\frac{d^2f(\lambda)}{d\lambda^2}\right|_{\lambda=\lambda_0}>0$$

$$\frac{df(\lambda)}{d\lambda}=0\Leftrightarrow -2(1-\lambda)R^2+2\lambda R^2=0$$

$$\Leftrightarrow \lambda_0=\frac{1}{2}$$

$$\frac{d^2f(\lambda)}{d\lambda^2}=4R^2>0$$

The last inequality is true for any typical value of R.

Hence, for $\lambda_0=\frac{1}{2}$ the minimum is obtained

Now we observe that there is a contradiction between the oscillatory form of the solution and the variance of $\underline{x}_2$. In section (5.18) we have shown that small values of λ (near 0%) create small oscillations while now we find that the minimum variance of $\underline{x}_2$

is obtained for a large value of λ (equal to 50%). So we have to find a compromise between them.

Now for $n \geq 3$ we shall develop the following analysis. Generally,

$$\mathrm{var}(\underline{x}_n) = \mathrm{var}(H_0 \underline{u}_{n-1} + H_1 \underline{u}_{n-2} + \ldots + H_{n-1} \underline{u}_0)$$

since the respective coordinates of each of the vector $\underline{u}_0, \underline{u}_1, \ldots, \underline{u}_{n-1}$ are independent variables we obtain

$$\mathrm{var}(\underline{x}_n) = \mathrm{var}(H_0 \underline{u}_{n-1}) + \ldots + \mathrm{var}(H_{n-1} \underline{u}_0) \quad \text{or}$$

$$\mathrm{var}(\underline{x}_n) = \sigma^2 \left\{ \mathrm{diag}\left(H_0 H_0^{(ts)}\right) + \ldots + \mathrm{diag}\left(H_{n-1} H_{n-1}^{(ts)}\right) \right\}^t$$

where diag(H) stands for the diagonios of matrix H

$H^{(ts)}$ stands for the transpose matrix of H.

As we can see from the last relationship for the variance of $\underline{x}_n$ it is very difficult to obtain the analytical form and consequently to apply the same procedure as previously for n=1,2.

5.20 Conclusion - A short review of the results

In this chapter, we have constructed a general model of multiple input - multiple output which is directly applicable to multinational pooling arrangement. The basic concepts are the interaction amongst the different subsystems (of each company) and the "harmonization action" i.e. the control action with respect to optimal interaction. The whole process targets to:

(1) "Smooth" as far as possible the surplus funds of each company participating in the pool (and consequently smooth the solvency requirement).

(2) Spread the experience (premiums - claims) of each company to the block of the other companies and obtain the desired solidarity amongst the insured groups of lives (as they actually constitute one large multinational company).

The specific modelling may be also used generally for the subsidiaries insurance companies of a parent company which aims to smooth the solvency requirement of each individual company or,

It may be used for capital allocation between different lines of business.

Closing this section (and simultaneously the chapter) we should mention briefly the important results.

(1) All versions I,II,III and IV of the general model have one equilibrium point equal to zero (given that $\varepsilon_1\neq0$, $\varepsilon_2\neq0$,..., $\varepsilon_m\neq0$ and $\varepsilon_1\neq0$, $\varepsilon_2\neq0$,..., $\varepsilon_m\neq0$). For model IV we have also obtained the required condition for stability (similar to the respective condition of chapter 4)

$$R-1<e\varepsilon<1 \ .$$

(2) All versions I,II,III and IV are completely controllable and observable so optimal control solutions may be designed.

(3) For models I and II a certain algorithm has been described in order to approximate a possible optimal solution with respect to the fastest response of the output. This optimal selection depends upon the choice of the parameters such that to obtain a root with the minimum possible magnitude

$$\rho_0=\sqrt[2m]{e_1\cdot\ldots\cdot e_m\varepsilon_1\cdot\ldots\cdot\varepsilon_m}$$

For model IV full investigation has been done with respect to fastest response and oscillatory form of the solution. It has been shown that small magnitude of the oscillation corresponds to small values of λ. Hence if we require a system with no large fluctuation we should choose small values for the interaction factor (λ) i.e. Each sub-system should be left "alone" (as possible) to arrange its problems.

(4) In section (5.19) it has been discussed optimality of Model IV with respect to minimization of variances in contrast with previous work. It has been shown that the optimal point is (for the variance of $\underline{x}_2$)

$$\lambda_0 = \frac{1}{2} .$$

Hence a contradiction appears (for the optimal selection of interaction factor) between less oscillations and smaller variances. The variance of $\underline{x}_1$ is independent of the choice of parameters (λ) while for $\underline{x}_n$ where $n \geq 3$ the calculation of the minimum requires numerical methods. (This restriction prevents the direct comparison of our result with that of Rantala (1987)).

Finally, we should stress that the general form of the model may be handled only using numerical methods, as the vectors and matrices involved are quite complicated!

Chapter 6

Application to the PAYG (Pay-As-You-Go) funding method

(An optimal control approach)

6.1 Introduction

In this chapter, we shall construct a model for the PAYG (pay-as-you-go) funding method using optimal control techniques attempting also to obtain a deeper insight into the demographics and the inertia mechanisms of the PAYG model. The final target will be the establisment of a new proposal which may comply with the respective socio-economic requirements (as inter-generational equity, solidarity, subsidiarity, people's expectations, government's planning etc.) and also applicable in the real world.

The PAYG method has been chosen as the subject of this research work because it is the basic vehicle (and in most times the only one) for funding the benefits paid by the Social Security Systems in many countries of the world. So we can say that (almost) everyone on this planet may be concerned (in some way) about the philosophy and the mechanisms underlying the process which is going to provide a respectable standard of living after retirement.

The concern mentioned above, is greater in the last years as the international situation (in demographical and economical factors) and the respective projections beyond the year 2000 have revealed funding problems threating the stability of the system or even the existence of it.

The structure of this chapter will be based upon the following steps:

- Describe the phenomenon of "aging populations".
- Discuss the philosophy and mechanisms of the PAYG funding method pointing also the future problems which currently threatens the system.
- Realize the current problem of the PAYG method.
- Motivate and explain the details of a new approach formulating the model using optimal control techniques.
- Solve the theoretical model considering three different approaches starting from the most general (and most difficult) and going down to the most practicable (providing an algorithm for applications)
- Apply the model into real statistical projection data of Greece up to the year of 2020 and obtain a "smooth" (consequently acceptable by all parts i.e. Government, Employers, Employees) path for contribution rates and age of eligibility for normal retirement.
- Draw conclusions, scope for further research and simulations of the model.

6.2 An international demographic trend ("Aging populations")

Most of the western developed countries (and not only) exhibits a certain demographic trend called "Population Aging" which according to the projections will be expanded rapidly after the year 2005. Although the title of the phenomenon is well descriptive, we shall quote Chen's (1987) definition who states that it is the "growth over time of the proportion of old persons according to some chronological age (usually 65) in the total population". So actually the population as a whole becomes more and more older!

But how does it happen? There are two basic reasons which the coexistence of them accelerate the phenomenon even more and guide the population structure to extreme patterns of high proportions of old aged lives.

The first reason is: The decrease in fertility rates. This decline, described as "baby bust" has followed the explosion of births which occurred immediately after the Second World War described as the "baby boom" of late 50's and 60's. Obviously, this certain decrease in fertility rates restricts the proportion of young people and sometimes (if it is below a certain level 2.1) also reduces the number of lives in the whole population (see Section (6.4) for further analysis of fertility).

The second reason is: The increase in life expectancy. Again this phenomenon is related to the big changes happened in Post War years and continued in this last decades. Of course, changes occurred in many areas of human activity, but here we refer to the great medical achievements which managed to enhance substantially the life expectancy. Now older people live even longer and consequently raise their proportion in the whole population.

At this point we must distinguish the two concepts of life expectancy (usually denoted in actuarial mathematics by e_x or $\mathring{e}_x$) and the limiting age (usually denoted by ω). Even the most modern medical achievements have not managed to affect (considerably) the limiting age of human lives (see section (6.4) for further analysis).

In order to obtain a "numerical taste" for the phenomenon described above, we shall quote an extract from the relevant table of Brown (1992) which presents an international comparison for "aging population" (see table (6.2.1)).

Table (6.2.1)

Aged Population Ratios (%)

Country	1985			2005			2025		
%	65+	75+	85+	65+	75+	85+	65+	75+	85+
UK	15.1	6.3	3.1	15.3	6.9	3.8	18.7	8.1	4.0
USA	12.0	4.9	2.6	13.1	6.7	4.1	19.5	8.5	4.8
China	5.1	1.4	0.5	7.4	2.4	1.0	12.8	4.1	1.8
India	4.3	1.1	0.4	6.1	1.8	0.7	9.7	3.1	1.3
Japan	10.0	3.7	1.7	16.5	6.4	3.0	20.3	8.0	4.9
⋮		⋮			⋮				

(Source: U.S. Department of Commerce 1987, 46-62). We shall also add some data for the Greek population (see next table (6.2.2).

Table (6.2.2)

Aged Population Ratio (%)

Country	1985			2005			2025		
%	65+	75+	85+	65+	75+	85+	65+	75+	85+
Greece	15.3	6.0	1.4	18.4	7.8	1.6	20.0	10.0	2.6

(Source: Greek National Statistical Service, Tables 1995).

Aged population ratio (65+) is defined as the ratio of the old people aged above 65, over the total population of lives.

As we observe the problem threatens more or less all the countries examined in the previous tables for the ages 65+ in all countries.

Apart from the absolute figures it will be interesting to consider the relevant increases between the extreme values of the tables. So we can see that for U.K. the increase in the population ratio 65+ between the year 1985 and 2025 is 23.8% while for India the same percentage is 97.9% i.e. the proportion of the population aged above 65 will be almost doubled.

From the analysis above we may conclude that some countries will exhibit bigger problems as the "aging" phenomenon develops rapidly.

Conclusively, "Aging Populations" is a reality. But is it really a bad situation? How does it affect the standard social structure with respect to the security system? How does it affect the PAYG model? These questions will be discussed in the next sections.

6.3 Description of the PAYG funding method

The PAYG (Pay-As-You-Go) funding method is the simplest one (at least in its primitive form) as it requires no accumulation of funds. Normally there is a small fund only for liquidity purposes (i.e. managing the incidence of cash flows). The whole insured group of lives is split into two subgroups: The active lives (or workers or contributors...) and the retired lives (or pensioners). A specific time period is also chosen, usually one calendar year, where the following equation holds:

$$\text{Total Contributions} = \text{Total benefits}\,, \qquad (6.3.1)$$

So actually, during the calendar year contributions are collected and benefits are paid while at the end of the year there is a requirement for zero balance between them.

The problem above, is carried forward year by year theoretically up to infinity. It is obvious that PAYG model is an unfunded system i.e. there is no accumulation of funds in order to support future benefits either for the existing active lives or the retired ones.

Each active life (in such a system) pays contributions to compensate the benefits of the retired lives while relying on the goodwill of the future generations of active lives who are going to contribute for his retirement benefits. But how willing will the next generation be to do so? Of course, this is the most critical question which actually touches the fundamental concept of this system i.e. Solidarity.

Inter-generational solidarity is the necessary requirement in order to build such a system. It is defined as the willingness of different groups of people (in the Social Security Context, the concept is defined amongst young and old generations) to participate in a common pool sharing actual experience (and consequently the loss (see Wilkie (1997)) of the environmental changes. This is not always easy and may be found when certain characteristics exist for the group of lives i.e.

1. Large number of lives (active, retired).
2. Great perspective for the continuous existence of the group theoretically (up to infinity).
3. The existence of a critical age where beyond that age the life can produce less than he needs for his living. (This should be true for the most of the individual lives).

The characteristics above may be easily realized in the whole population of a certain country.

Apart from the solidarity concept, there are two other important ones: the inter-generational equity and subsidiarity which are actually complementary.

Inter-generational equity is defined as the situation where, under a certain type of measurement (normally financial one) all generations are equal to each other. (The figure for the measurement is the same for each generation).

The equity concept is the great demand of each society. So a good and clear process for its measurement should be established. The most usual one, is the calculation of the implied rate of return. Quoting from Lapkoff (1991) we obtain that "The rate of return is the interest rate that equalizes the stream of contributions to the stream of benefits and would have been the interest rate applicable had the contributions actually been invested". This rate of return is calculated either for individual or cohorts of lives. The "rate of return" is equivalent to the "Internal Rate of Return IRR" (see McCutcheon & Scott (1986)). The comparison between the implicit rates of return reveals the existence and the magnitude of the equity concept of the system.

The subsidiarity concept complements equity as it is impossible to measure everything with figures or avoid random events which sometimes result to catastrophes. Subsidiarity is defined as the willingness of different groups of people to abandon the equity concept (i.e. partially abandon their rights) and contribute in favour of other groups of people. So, for example a generation who lives during a war may not contribute too much in the social retirement scheme but surely must receive pensions at the retirement age. The lives of the next generation should accept to subsidise them (not only because they have fight in the war) as being their fathers or grand fathers. Conclusively, we could say that subsidiarity is the extreme form of solidarity (Willekes and van den Hoogen (1998)).

Closing this section we should mention that apart the inter-generational solidarity, there are also other forms of solidarity between (as quoted in Willikes and van den Hoogen (1998)):

- people with high and low income (income solidarity)
- males and females
- (unmarried and married) people who cohabit with single people
- households with one income and households with a double income
- healthy and less healthy people

6.4 The demographics of PAYG model

In this section, we shall discuss the different demographic factors and the way affecting the PAYG model i.e.

(1) Fertility

(2) Immigration (Positive)

(3) Labour Force Participation

(4) Mortality

(5) Morbidity

(6) Withdrawal or Early Retirement

Generally speaking, the factors may be separated into two categories. The first affects entrance to the system ((1), (2), and (3)) while the second affects exit from the system ((4), (5) and (6). Let's analyze separately them one by one.

(1) Fertility: This is the most important factor as regards the entrance of the system. Before we go further we shall define the concept of the "replacement ratio" which actually equals 2.1. "Replacement Ratio (RR)" is the required fertility rate (births per female member) in order to replace the population generations (Brown (1992)) i.e. obtain a stationary population. So the 2nd condition of section (6.3) which requires this continuous existence of the group of lives may hold for ever as the population will be increasing or at least constant. In the opposite case where the fertility rate is below the (RR) then the population is decreasing in the future and the PAYG system will exhibit a catastrophe.

It is also important to examine not only the absolute value of the fertility rates but also the trend between consecutive values. Increasing fertility rates will result decreasing contribution rates for the PAYG system while a decline in fertility rates (as it is the current situation, described in section (6.2)) will result in relevant increases.

Hence, the government (or the responsible authority for the smooth operation of the Social Security System) should have some kind of control mechanism for birth rates (if possible). For example a control mechanism may be established providing financial incentives (additional income or tax relief) for families with three or more children. At this point we should mention two of the important schools of thought for the time development of fertility rates which have been described in recent years (we shall quote Brown (1992) for the relevant description).

"One school is represented by Easterlin "wave theory" of fertility (Easterlin, 1987). Easterlin postulates that fertility rates rise and fall in a wave-like pattern with a cycle length (peak to peak or trough to trough) of two generations".

"The other common theory on fertility is presented by Ermisch (1983) Butz and Ward (1979) and others. The theory states that in a one-earner family if the worker's real wages rise rapidly and the cost of children remains constant then family will have more children... In a two earner family if real wages rise rapidly but the wife has to leave the work force or interupt a career path to have and raise children then the cost of children rises and fertility rates will not change"

(2) Immigration: This factor actually expands the work force of a country (if positive) or reduce it (if negative). So the whole analysis as regards the positive immigration. So the whole analysis may be parallel with fertility except that the migrants are aged x which is not necessarily zero. Of course births are all aged zero!

The quality of the workers (skilled or not, educated or not,...) will be a critical issue for the control which should be established by the Government. Increasing entrance rates of unskilled workers may result bigger problems in Social Security System as training cost, unemployment benefits etc will offset any advantage of their contributions.

(3) Labour Force Participation: A secondary but also important index is the level of participation in the labour force of the potential active population. Obviously higher participation rates result in lower contribution rates for the PAYG system. So if a government cannot successfully control the births, it may attempt to provide incentives in order to increase the labour force participation rate and consequently lessen the contribution rates for the Social Security System.

(4) Mortality: The important factor as regards exit from the system is the pattern for the mortality rates especially those of the retired lives (which are usually higher than for active lives). Lower mortality rates will result an enhanced life expectancy and consequently more retired lives with larger benefits as a total and finally higher contribution rates for the existing active lives. (This is the current demographic trend described in section (6.2).) In the opposite case where there is an increasing trend for the mortality pattern and hence lower life expectancy, the figures go the other way round.

We should stress again that medical achievements have managed to shift the life expectancy by keeping in older age bands more and more humans and not by shifting considerably the limiting age which remains almost the same during the last decades. Of course, enhanced life expectancy have considerable affects also in health care treatment raising again the cost of the Social Security System.

Now as regards the controllability of this factor the only way of thinking may be an improvement. No government will attempt to increase mortality rates in order to reduce the cost of pensioners and finally relieve the Social Security System!!!

(5) Morbidity: Morbidity pattern is basically applicable to the active lives. Low morbidity rates will secure a satisfactory number of workers and consequently acceptable levels of contribution rates for Social Security System. Morbidity pattern is also applicable to retired persons in conjunction with other recent changes in Society (e.g. the family structure) may increase the cost of Social Security System for the long term care of disabled retired persons or the respective disability pension for them.

(6) Withdrawal or early retirement: Similarly with the labour force participation rate, the withdrawal rate defines the people remaining in active lives. Small withdrawal rates will result in (finally) smaller contribution rates. This is a factor where the Government may intervene and encourage late retirement (so actually low withdrawal rates) and hence keeping more lives as active workers.

Closing this section we shall refer to the normal retirement age which one of the most important issues for the calculation of the contribution rate. The normal retirement age is a powerful control variable in Government's hands which may help to reduce the cost of Social Security System. Of course any decision should take in account people's reactions and potential acceptance by the society.

Obviously, the normal retirement age draws the line separating the cohort of active lives from the cohort of retired persons. Shifting this line to the right (increase of retirement age) will result in a smaller cohort of retired persons compared with the larger cohort of active lives and ultimately smaller contribution rates. The opposite reasoning may be applicable if we decrease the retirement age.

Finally, we should also mention the entry age as another control variable of the system. Of course age at entry may be fully controllable in a private pension scheme but not in the social security system.

6.5 The economics of PAYG model

The other critical issue of the PAYG model is the economics of the system. Returning back to the basic equation (6.3.1) which balances the total contributions and benefits in each time period (one year) we should stress that (see Brown (1992)):

(a) Contributions are related to wages and salaries which move in line with the economic growth and productivity while

(b) Post-retirement benefits are related to price inflation (securing a certain standard of living for retired persons).

So there is a gap between the rates which affect the contributions and benefits respectively. Normally this is a positive gap (since the economic growth normally exceeds the price inflation) i.e. we have a positive real economic growth. Now the bigger the gap is the lower contribution rates will be applicable in the future.

Obviously, the gap may be fully controllable by each Government as the second item of the increase of benefits is directly defined by Government's decision. So it may be a case where a decision may be taken for lower increases (than price inflation) in order to increase the gap and finally decrease future contribution rates. That means, the retired persons will pay some of the cost of the Social Security System by losing the initial standard of living.

Of course, such a decision is painful for a Government as the people won't accept this solution and many social problems may occur.

The expenditure for the Social Security System should also take into account not only the retired persons (so the elderly dependency ratio) but also the training cost for the young ages or the unemployment benefits. It should be considered in the wider context of the Government's budget which is the ultimate object which should be balanced during a certain time period.

In the economics of PAYG we should also place the important implicit rate of return which defined in section (6.3) and according to Samuelson (1958) equals to the population growth plus the real wage growth rate (further analysis for this item will be provided in the next section).

Another issue in the economics of the PAYG model is the non-existence of an accumulation fund (or a small fund for liquidity purposes). Normally the equation (6.3.1) stands alone with no provision for funding future benefits. In practice, modern models (see Nesbitt et al (1995)) assume some kind of a stabilization item in equation (6.3.1) in order to manipulate in a more efficient way the whole system. In section (6.8) we examine how the existence of a stabilization fund may improve the total performance of the system and suggest a solution with the existence of a contingency fund.

Finally we must stress that the PAYG funding method may be used in the days of high inflation as the other fully funded methods exhibit difficult problems by inflation. (Inflation has destroyed several fully-funded schemes in Europe this century (see Trowbridge (1977)).

6.6 The basic structure of PAYG model (Defined Benefit & Defined Contribution Plans)

As we generally know, there are two categories of pension funds: those with fixed contribution for their members (Defined Contribution) and the others with the fixed benefit (Defined Benefit). This categorization may also be applicable to the plans operating under a PAYG model. In order to investigate the mechanisms of the two categories mentioned above, we rewrite equation (6.3.1) while expanding it in the following form.

$$(A.L) \times (I.C) = (R.L) \times (I.P) \qquad (6.6.1)$$

where,

A.L	=	total number of active lives
I.C	=	individual contribution
R.L	=	total number of retired lives
I.P	=	individual pension

So there are two options either to fix (I.C) or (I.P) in order to produce a certain type of plan. Now from equation (6.6.1) we may derive that the size of the cohort of lives whether for active lives or retired persons determines the relevant size of benefits or contributions.

For a defined benefit plan, large cohorts will have smaller contributions while smaller cohorts will exhibit larger contributions. For a defined contribution plan the situation above is applicable for benefits i.e. large cohorts obtain small benefits while small cohorts receive larger benefits.

A very interesting research for the advantages & disadvantages of the two types of plans is that of Lapkoff (1991) who also refers to Keyfitz (1985). The two authors investigate which is the most efficient way (fixing contributions or benefits) to obtain inter-generational equity. The last condition is translated into the requirement of equal rates of return (as defined in section (6.3)).

As Lapkoff (1991) states in a stable population, ignoring the real economic growth (consequently and in a stationary one) the PAYG model produces identical rates of return for every cohort of lives (equal to the growth rate of the population). The problems arise in a non-stable population with cohorts of lives of different sizes. In this case, the large cohort of lives obtain high rates of return under a defined benefit plan, since the benefit is fixed while the contribution for them is smaller and the small cohorts of lives obtain greater rates of return under a defined contribution plan (using similar reasoning as before).

Finally, the problem (as investigated by Keyfitz and Lapkoff) can be expressed as a minimization problem of the standard deviation of the different obtained rates of return for each cohort of lives.

Keyfitz's simulations showed that defined contribution plans produce rates of return (for the different cohort of lives) with a smaller standard deviation than the respective defined benefit plans. As Lapkoff points, Keyfitz's result can not be generalized as he used an extract of a certain demographic pattern (The Post War baby-boom) ignoring the previous historical data. Furthermore, Lapkoff continued the investigation using a theoretical model of a stationary population with four cohort of lives (three of them as active and one for retired persons). Then the specific population was subjected to one abnormally sized cohort of lives.

Calculating the rates of return under a defined benefit and a defined contribution structure she proved that the minimum standard deviation is obtained for the defined benefit plan. Hence, it is more fair and consequently should be used for a PAYG model.

In the real world most of the Social Security Systems have adopted the defined benefit structure or sometimes a hybrid model (a combination of the two structures) but which is heavily-weighted towards the defined benefit structure.

In Section (6.9), the basic structure of Lapkoff's model will be used as a vehicle to illustrate how the existence of a contingency fund in the PAYG method may secure (in a better way) the intergenerational equity by producing small deviations for the different obtained rates of return for each cohort of lives.

We shall return to the concept of the fund in section (6.9) providing the motivation for our proposal of a well-operated PAYG model.

6.7 The existing problem of Social Security Systems operating under the PAYG model

We have described up to now all the basic features of the underlying philosophy and inertia mechanisms of the PAYG funding method. Now if we combine this literature with the existing international demographic trend of section (6.2) we may easily identify the problem which actually threatens with an absolute catastrophe the Social Security Systems of all countries operating under the PAYG model.

We must also point out that the two demographic trends are placed in the same side so the ultimate additive effect accelerates the whole system even more. Decline in fertility rates and enhanced life expectancy with even worse projections (the projected elderly dependency ratio equals to 2:1 the year 2040 for the countries of the European Union) for their development, force the contribution rates up to unacceptable or even illogical levels for the future generations of active lives. The potential catastrophe is not just a theoretical threat but it describes the possible real situation i.e. It is very possible that generations beyond the year of 2010 or 2020 won't accept to meet the cost of the baby boom of late 50's and 60's as the active cohort of lives will be quite small in order to subsidise the cohort of retirees.

Now the critical question which should be answered is the following:
"What can we do in the future in order to avoid undesirable problems with the funding procedure of the Social Security System?"

A lot of individuals and organizations around the world have attempted to solve the problem and propose possible actions in order to avoid the catastrophe of the system.

The World Bank (1994) in its book with the title "Averting the Old-Age Crisis" proposed a multipillar approach. The first pillar will be supported by Social Security providing only a minimum guaranteed income. The second pillar will be supported by private fund management (insurance companies, pension funds etc...) providing a pension directly related with the contribution paid of each individual. Such a model has already been established in Chile from 1981 producing quite acceptable results up to now. Daykin (1998) provides an extensive review of the Chile's model and all the other experience from different countries around the world which tried to reform their Social Security System.

Of course this second pillar proposal is not a panacea but it is a wise action towards the right action. And as Daykin (1998) states "We can't prevent the ageing of population but we can at least try to reduce the risk of it becoming a crisis".

Now in order to reduce the risk of crisis and produce a balanced Social Security System, there are some minor and major control variables. The minor variables are (a) the method of indexation of pension payments (b) the pattern for the accrual rates of benefit (c) the period of averaging of the final salary.

The major two control variables in order to prevent sustainability of the system is the contribution rate and the normal retirement age.

Of course the control and consequently the modification of these variables is not an easy exercise as any solution or action should be acceptable by the society and the lives which are going to pay the cost. In order to achieve this target actuaries should communicate efficiently with all the parties (Government, Employers, Employees) identify the expectations of each one and finally propose a suitable solution which will be finalized by political terms and conditions.

6.8 Different solutions proposed for the problem of social security system

In this section, we shall focus on three approaches (the second one actually consists of a series of three papers) which discuss the three major control variables of the PAYG model. They are very interesting as all the authors (of the respective papers) use an entirely different approach from the other and consequently the whole puzzle of the PAYG model and Social Security System is revealed. Finally we should stress that, this is not a literature review but a selective presentation of solution proposals.

The three approaches are presented below providing a small summary and some critical comments.

1) **Vanderbroek (1990): "Pension Funding and Optimal Control.**

   The paper formulates the PAYG model using optimal control techniques. It focus on the contribution rates searching for the design of an optimal path. The basic criterion for optimality relies on the smoothing procedure of contribution rates along with the smoothing of the fund values. This last requirement (i.e. the smoothing of fund values) may not comply with our suggestion as appears in section (6.9). As we are going to see in section (6.9) the potential ability of the fund to fluctuate deliberately may absorb random effects and so improve the performance of our model. Hence, it should be interesting to examine a similar model eliminating the requirement of a smooth path for fund values.

2) **(i) Nesbitt (1991), "Elementary models of reserve fund for OASDI in the USA.**

   (ii) Nesbitt et al (1995), "Some Financing Options for Social Security"

   (iii) Nesbitt et al (1995), "Conclusions from Michigan Studies of Social Security Financing".

The papers propose the introduction of a contingency fund into the standard PAYG model in order to avoid large fluctuations in equilibrium rates. It discusses the maintenance of such a fund up to the levels of "100 to 150 percent of the current's year outgo for benefits and administration".

It also introduces the concept of the n-year roll forward reserve (where $n \geq 2$). This concept may stand between a full prefunding method and the PAYG model. Actually if n goes to infinity (or the limiting age of the current group) the n-year roll-forward reserve is a full-funding method. The n-year roll-forward reserve is defined as the sufficient reserve for covering liabilities and administration expenses for the n future years.

So each year annual contribution will provide the projected benefits n years ahead rather than the outgo of the current year. The maintenance of such a contingency fund will prevent large fluctuations in the contribution rates.

Finally, we shall provide some basic algebra for the "n-year roll forward reserve method". The required notation is the following: (see Nesbit et al (1995))

${}_nA_{31/12/k}$ = Present value of the outgoes of the next n years at 31/12/k.

O_k = Outgo during the year k (occurring in the middle of the year).

${}_nI_k$ = Required contribution in year k (occurring in the middle of the year) in order to cover the projected outgo n years ahead.

δ = constant force of interest.

Then we obtain the relationships below:

$${}_nA_{31/12/k} = O_{k+1}e^{-\frac{\delta}{2}} + O_{k+2}e^{-\frac{3\delta}{2}} + \ldots + O_{k+n}e^{-\frac{(2n-1)\delta}{2}} \tag{6.8.1}$$

$${}_nA_{31/12/k}e^{\delta} + \left[{}_nI_{k+1} - O_{k+1}\right]e^{\frac{\delta}{2}} = {}_nA_{31/12/k+1} \tag{6.8.2}$$

$${}_nI_{k+1} = O_{k+1+n} \cdot e^{-n\delta} \tag{6.8.3}$$

There is no smoothing procedure used in this model.

3) **Brown (1992): Pay-as-you-go funding stability: An age of eligibility model.**

The paper presents a full description of the demographics of the PAYG model. It also places the problem of Social Security into a wider context of the Government's budget and planning for Canada. It states the current demographic trends and the respective problems of Social Security Systems. Finally, the author proposes a solution (which may be easily adopted by the Canadian Government) based on the control (increase) of the normal retirement age while keeping constant the total expenditure of the Government for youths, unemployed and retired persons. He actually applies his ideas to the Social Security System of Canada defining the index LFEDR (Labour Force Expenditure Dependency Ratio), where

$$\text{LFEDR} = \frac{(1.7 \times Y) + (1 \times U) + (4.244 \times O_{65+k})}{LF} \tag{6.8.4}$$

Y = those aged 0-19

U = unemployed labour force

O_{65+k} = aged 65+k and over

LF = projected labour force

So keeping constant the LFDER index beyond the year 2006 he proposes a gradual increase of the normal retirement age up to 2030.

6.9 Discussion and Motivation for a new proposal based on the concept of the "Intergenerational Equity"

We have seen up to now the basic features of a PAYG model (i.e. demographic and economic factors, potential structures, current problems caused in Social Security Systems and the respective proposed solutions). At this point, we shall attempt to use all this experience and data in order to formulate a new proposal which will be heavily based on the concept of "Intergenerational Equity".

As we have seen in the three approaches of section (6.8) the three major variables involved in the PAYG model are:

1) The reserve (surplus / deficit) fund.
2) The age of normal retirement.
3) The contribution rate.

Of course there is another one, the relevant pension benefit (i.e. the entitlement formula or the rate of benefit increase each year etc.).

Keeping in mind the analysis in section (6.6) and the results of Lapkoff which is in favour for a defined benefit plan, we shall construct such a plan (which corresponds better to "Intergenerational Equity" by

a) Leave constant the entitlement formula and

b) Assume also that Government always decides that benefit increases should be equal to the annual inflation rate, so actually preserving the standard of living for the retired lives.

Now, returning back to the three major variables let's proceed with some further analysis and the proposal for the role of each of them into our PAYG model.

1) The reserve (surplus / deficit) fund

That usually, equals to zero (for the standard PAYG model). But as we are going to prove with our proposed modification, the existence of a certain contingency fund may improve the performance of the system with respect to "inter-generational equity". We shall assume a ficticious demographic pattern in our approach similar with that of Lapkoff (1991) quoting also from him the first five (5) assumptions as below. The sixth assumption contains our proposal for the introduction of a contingency fund. Hence we have the following assumptions.

(1) "A population with four age groups - three of working population and one of re-tired".

(2) "To keep the arithmetic simple the population is assumed stationary with one individual at each age group and no mortality until the end of retirement".

(3) "Each worker (of the active population) contributes \$1 per time period and thus \$3 over their working lives".

(4) "In any one time period \$3 is collected altogether and distributed to the retiree". So equation (6.6.1) always holds during one time period.

(5) In a certain time period, say t, "the stationary population is subjected to one abnormally-sized cohort" of a double size. (For simplification we shall assume that $t^{*}=2$).

(6) After the occurrence of the abnormally-sized cohort and for the time periods (t=2,3,4,5) which remains in the active (t=2,3,4) or retired (t=5) population of lives, equation (6.6.1) does not hold while a contingency fund is created since the contributions and benefits remain constant (per person and per unit time). The fund exists at times t=2,3,4 while it disappears at the end of the 5^{th} time period compensating the pension benefits of the abnormally-sized cohort of lives.

Under the situation described above, the rate of return for all cohort of lives (even for the large one) will be the same (in the specific example equals zero which is the actual growth of a stationary population) (see table (6.9.1)).

Another situation may not be so extreme i.e. at time t=2 reduce the contribution as the large cohort appears but not down to 0.75 units, perhaps down to 0.9 units so a surplus of 0.6 units may be accumulated. Applying the same rule at times t=3,4 we may obtain an accumulated surplus of 1.8 units at time t=4 which will partially compensate the excess benefits for the retirement of the large group. At that time t=5 contribution for active lives will be raised to 1.4 units. Even in this situation the produced rates of return will have a smoother pattern than Lapkoff's original model but not zero deviation as they exhibit in the extreme previous case where the contributions remained stable (see table (6.9.2.)).

Finally, we may conclude that the existence of a "contingency fund" will have a favourable effect to the PAYG model as it:

(1) Absorbs the random fluctuations in the different demographic patterns.

(2) Smooths the produced rates of return for each cohort of lives whether small or medium or large ones.

(3) Ultimately, approximates in a better way the concept of inter-generational equity by minimizing (as far as possible) the standard deviation of the produced rates of return of each cohort of lives.

So we shall use a non-zero reserve fund in our model, which has two characteristics.

(a) It fluctuates deliberately (in the short run) in order to absorb random fluctuations in mortality or fertility rates or due to any other cause.

(b) It returns to zero when the randomness disappears leaving the system to a new equilibrium point with respect to the other two variables.

In order to obtain (a) and (b) we should build a control model in where the reserve fund "should have the ability" to "distinguish" the random or the constant nature of a certain demographic pattern. Such a "clever" system may understand that the demographic pattern in Lapkoff's model is random and fully absorbs the fluctuations leaving constant the contribution rates and age of normal retirement (while in opposite cases fully pass the event to the other two variables).

Table (6.9.1)

Time	Number of Active lives	Number of Retired lives	Total Contributions	Total Benefits	Annual Balance	Accumulated Fund
1	3	1	3	3	0	0
2	4	1	4	3	1	1
3	4	1	4	3	1	2
4	4	1	4	3	1	3
5	3	2	3	6	-3	0
6	3	1	3	3	0	0

Table (6.9.2)

Time	Number of Active lives	Number of Retired lives	Total Contributions	Total Benefits	Annual Balance	Accumulated Fund
1	3	1	3	3	0	0
2	4	1	3.6	3	0.6	0.6
3	4	1	3.6	3	0.6	1.2
4	4	1	3.6	3	0.8	1.8
5	3	2	4.2	6	-1.8	0
6	3	1	3	3	0	0

2) The age of normal retirement

The age of normal retirement, determines the actual number of active and retired lives and consequently the potential wages / salaries and pensions. This variable should be linked with the concept of life expectancy. It may be fair that all cohorts of lives should rest (i.e. live as retired persons) equal percentage of their total lifetime. That means cohorts with high life expectancy should retire at a higher age and cohorts with low life expectancy to retire earlier. This may also comply with the view of Dilnot et al (1994) who points that "If the population is aging because of increased longevity then individuals will need a longer period in the labour force to obtain a given level of average consumption over their lifetime. This might lead to individuals prolonging their working lives by postponing retirement...". Keeping the rule of "equal percentage of rest" we may produce the following example. Assume that currently exists a cohort of lives with life expectancy (at age zero i.e. $\mathring{e}_0$) equal to 80 and the current age of normal retirement is 60. Then if another cohort of lives appears with life expectancy equal to 84 then the age of normal retirement should be raised to 63 (the equal percentages may also be applicable but considering not the life expectancy but the years of contributions and the years of receiving benefits from the system).

3) The contribution rate

Finally, the contribution rate should be linked with the fertility rates and the respective population growth rate. The higher the growth rate the smaller the contribution rate which should be applied to the PAYG model. A certain rule (similar to the one found for the age of normal retirement and life expectancy) may be described in order to link and control the contribution rate with respect to the changes in the population growth rate. Of course the PAYG model should be able to identify whether it is a constant or random change in the population growth rate and act accordingly.

Having established the underlying philosophy of the three major variables we shall construct a "clever" control model using them.

The reserve fund will fluctuate deliberately absorbing random demographic patterns while targeting the zero value at certain terminal time points.

The age of normal retirement and the contribution rate will be controlled guiding them through a smooth path over time.

The smoothness of the path will be determined by a functional which shall weight the changes in the two variables. The weights will be parameters determined firstly by the people's expectations and secondly (a) from life expectancy for the age of normal retirement and (b) from the population growth rate for the contribution rate. People's expectation is a very important parameter which should be examined by statistical research, be interpreted and finally be incorporated in our model, modifying the weights (which have been calculated by technical reasoning before) of the two control variables i.e. In the previous example with respect to the life expectancy and the age of normal retirement the new cohort of lives with the high life expectancy (equal to 84) may prefer to pay a slightly higher contribution rate and ultimately retire also at age of

60. Hence the weights in the smoothing process will be heavily based on people's expectations.

6.10 Formulation and notation of the proposed model using optimal control techniques

In this section we shall proceed with the typical formulation of our proposed model interpreting the general discussion and motivation of the last section into symbols and equations. The proposed model is a deterministic one (but it may easily be transformed to a stochastic one).

The problem will be defined in the continuous form and in the next sections we shall degrade it, into the discrete type.

Firstly, we shall define the symbols

F(t) : Reserve (accumulated) fund at time t

c(t) : Contribution rate at time t

r(t) : Normal Retirement Age at time t

pl(t,x) : the population at time t aged x

s(t,x) : total salary received by a person aged x at time t.

b(t,x) : total benefit paid to a life aged x at time t.

a : entry age in the labour force.

ω : limiting age of the life band ($l_\omega=0$)

B(t,x) : Total benefits to be paid at time t if the relevant retirement age has been fixed to age x i.e. $B(t,x)=\int_a^x pl(t,y)\cdot s(t,y)dy$

W(t,x) : Total wages / salaries (where the contribution rate c(t) is applicable) at time t if the relevant retirement age has been fixed to age x i.e.

$$W(t,x)=\int_x^w pl(t,y)\cdot b(t,y)dy$$

(We assume that functions B(t,x) and W(t,x) for benefits and wages are known for every time t i.e. $t\in[0,\infty]$ and for any age in the life-band i.e. $x\in(0,\omega)$. These forms may be derived from projection data of the demographics of the population).

c_0 : A standard value for the contribution rate. We may consider it as an initial value or average value near which, we try to place the path of the future consecutive values of C(t) using the smoothing process.

r_0 : (Similarly with c_0). It is the standard value for the retirement age.

δ : The constant force of investment rate of return applicable to the reserve fund.

θ : The weight applicable to any change of the contribution rate from the standard value c_0 (consequently, $1-\theta$ is the weight applicable to any change of the retirement age from the standard age r_0).

Having defined the symbols above we may proceed with the formulation of the equations. The first one describes the development of F(t) i.e.

$$F'(t)=\delta F(t)+c(t)W(t,r(t))-B(t,r(t)) \tag{6.10.1}$$

That is a differential equation for F which may be written generally as

$$F'(t)=G(F(t),c(t),r(t)) \tag{6.10.2}$$

where G is a non-linear function of F(t),c(t) and r(t). As we can see equation (6.10.2) determines a dynamic system where

F(t) is the state variable and c(t), r(t) are the control (input) variables.

The c(t) and r(t) are the potential control variables for a Government when trying to balance the Social Security System. Of course there is another one the increase in pension benefits but as we have stated this variable should not be considered as a control variable assuming that it is fixed or linked with the relevant annual inflation rate.

Now the objective for the authority responsible for a Social Security System is the determination of a "smooth" path for the control variables guiding the system over time and targeting a zero (or almost zero) fund value. The existence of a "smooth" path requires that the relevant choices for c(t) and r(t) will comply with people's expectations and the other technical criteria described in section (6.9).

Summing up all the literature above the question of optimal path may be answered by minimizing the following expression i.e.

$$\min_{c(t),r(t)} \int_0^T \left\{\theta\cdot\left[100\cdot\left(c(t)-c_0\right)\right]^2+(1-\theta)\cdot\left(r(t)-r_0\right)^2\right\}dt \qquad (6.10.3)$$

The coefficient of 100, has been applied in the change of contribution rates in order to arrange the metric problems which exist as c(t) is a percentage less than unity and r(t) is a number greater than unity (near 65).

The weights θ and $1-\theta$ reveal the negative effect which occurs upon a change in the control variables c(t) and r(t) respectively. The parameter θ would be obtained after statistical research and negotiations with all parties involved in the Social Security System (i.e. Government, Employers, Employees etc...).

Hence expression (6.10.3) minimizes the negative effects in the Society upon the certain changes induced in the Social Security System with respect to contribution rates and age of eligibility.

As we have seen in section (3.11), O' Brien (1987) uses a similar approach smoothing the fund level around the desired level defined by the funding ratio.

Vanderbroek (1990) also proposes a similar expression for minimization with respect to the contribution rates and (instead of the age of normal retirement) the fund level.

Haberman and Sung (1994) proposes the minimization of the control error and/or the control action error i.e. minimization of the differences or the square differences of the actual and desired solvency levels and contribution rates respectively.

The functional of expression (6.10.3) does not contain the fund values as we assumed that will fluctuate deliberately in order to absorb random events. Of course we shall require a small value for the fund level at the end of the respective period of examination at time t=T

i.e. $$F(T) \in (-\gamma, +\gamma),\ \gamma > 0 \tag{6.10.4}$$

or more strictly $F(T)=0$. (6.10.5)

Combining equations (6.10.1) (6.10.3) and either (6.10.5) we obtain the typical form of an optimal control problem as below.

$$\left.\begin{array}{c} \min\limits_{c(t),r(t)} \int_0^T \left\{\theta\left[100\cdot\left(c(t)-c_0\right)\right]^2 + (1-\theta)\cdot\left(r(t)-r_0\right)^2\right\}dt \\ \\ F'(t) = \delta F(t) + c(t)W(t,r(t)) - B(t,r(t)) \\ \\ F(0) = F_0 \quad \text{and} \quad F(T) = 0 \end{array}\right| \tag{6.10.6}$$

In the next sections from (6.13) up to (6.17) we shall develop three different options of this basic system. The first called [C-C] model will be exactly the same with that described in system (6.10.6) which has continuous functions for Wages (W) and Benefits (B) while the decision for contribution rates and age of eligibility is also taken continuously (i.e. c(t) and r(t) are control variables of a continuous form).

The second version of the model called [C-D] will have continuous functions for W and B but discrete form for c(t) and r(t) i.e. the decision for the control variables will be taken at discrete points of time.

Finally the third version of our model called [D-D] will consider discrete forms for all functions while we are going to relax the third strict condition of (6.10.5) substituting with condition (6.10.4) i.e. we shall demand to guide the fund value of the system to be "near" zero (and not exactly at the zero point).

6.11 General discussion applying the model to a stationary, stable, non-stable and stochastic population

It will be interesting to investigate the behaviour of our model under different demographic population patterns in order to obtain a further insight for its mechanisms, solution and finally the relevant level of efficiency for funding a Social Security System. For this purpose, we shall examine four different situations as stated in the title of this section (i.e. stationary, stable, non-stable and stochastic populations) assuming that

$$s(t,x)=s \text{ and } b(t,x)=b \quad \forall t,x \tag{6.11.1}$$

(i.e. salaries and benefits are fixed over time and age). This simplification does not destroy the generality of our results but facilitates our calculations and focus our attention on the specific demographic pattern.

Stationary Population

Stationarity is not a realistic assumption for an ageing population but we shall also examine this case as provides a good framework (i.e. a simple population pattern) in order to build gradually our insight to the model and understand the interaction between the control variables c_n, r_n, F_n.

So in the simplest case, of a stationary population the number of lives at each age band is constant every year, so the functions of wages W(t,x) and benefits B(t,x) do not depend on the time variable t but only on the age variable x. All the basic indices are also constant i.e. fertility rates (population growth equals zero) life expectancy etc. So at the beginning of such a PAYG system we may choose the optimal pair of values (contribution rate and age of eligibility) which complies with people's expectation using the formula below

$$c_0 \cdot W(t,r_0) = B(t,r_0)$$

or analyzing W and B we obtain

$$c_0 \cdot s \cdot \int_a^{r_0} l_y dy = b \int_{r_0}^{\omega} l_y dy \tag{6.11.2}$$

(where l_y is the usual function of a life table, i.e. the number of lives at age y)

where (c_0, r_0) is the optimal pair. Now according to the value of the initial fund F_0 we have the following cases:

1st Case" $F_0=0$

Then $\quad c(t)=c_0$ and $r(t)=r_0\ \forall t \geq 0$ (6.11.3)

is the optimal path for c(t) and r(t) which satisfies the system (6.10.6) i.e.

(1) The expression (6.10.3) is minimized under (6.11.3) and the respective minimum value is zero.

(2) Equation (6.10.1) with the respective condition (6.10.5) also holds for the trivial function F when

$$F(t) = 0 \qquad \forall t \geq 0$$

If we sketch the Fund value F(t), the contribution rate c(t) and the age of eligibility r(t) with respect to the time variable t (in a four dimension linear space) then the produced graph will be a straight line parallel to the axis of the time variable and starting from the point $(F,c,r,t)=(0,c_0,r_0,0)$

2^{nd} Case: $F_0 \neq 0$

Under this case we must distinguish two possible scenarios

(i) The fund value F_0 is a small one such that condition (6.10.4) is satisfied then we may follow the solution of the first case above

i.e. $c(t)=c_0$ and $r(t)=r_0$ $\forall t \geq 0$

while F(t) remaining constant equal to F_0.

(ii) The fund value F_0 is <u>not</u> a small one, so we require to increase or reduce it up/down to a certain level in order to satisfy condition (6.10.4). Then we should choose initial values for the contribution rate and the age of eligibility slightly greater or smaller than c_0 and r_0 in order to increase or reduce the value of F_0 (if positive or negative respectively).

As the time passes F(t) will go near to zero and the c(t), r(t) near to their optimal values of c_0 and r_0 respectively i.e.

$$\lim_{t \to \infty} c(t) = c_0 \quad \text{and} \quad \lim_{t \to \infty} r(t) = r_0 \ . \tag{6.11.4}$$

Hence the graph of the optimal path will be an asymptotic line (which may be the same after a certain point) to the straight line described in the first case. It is obvious that a compromise should be achieved between the starting point and the slope of the optimal path in order to minimize the first condition of the system (6.10.6). For example we may choose c(0) and r(0) such that

$$F(0) = B(r(0)) - C(0) \cdot W(r(0))$$

So the fund F will become immediately zero and then choose

$$c(t)=c_0 \text{ and } r(t)=r_0 \qquad \forall t > 0$$

Stable population

The stable population may be considered as a generalization of the stationary as the number of lives is not constant as time passes but increases or decreases with a constant rate, say g (i.e. stationary is a stable population with g=0). A negative g would imply "ageing".

Again we can choose a pair for the contribution rate and retirement age i.e. c_0, r_0 respectively. Then we may write the formula

$$c_0 \cdot W(t, r_0) = B(t, r_0) \qquad \text{or equivalently}$$

$$c_0 \cdot \int_a^{r_0} pl(t, y) \cdot s dy = \int_{r_0}^{\omega} pl(t, y) \cdot b dy \qquad \text{or}$$

$$c_0 \cdot s \cdot \int_a^{r_0} l_y \cdot e^{-gy} dy = b \int_{r_0}^{\omega} l_y \cdot e^{-gy} dy \tag{6.11.5}$$

and if we consider g as force of interest we may rewrite equation (6.11.5) in terms of the continuous annuity values i.e.

$$c_0 s \cdot \bar{a}^{(g\%)}_{a:\overline{r_0 - a|}} = b \bar{a}^{(g\%)}_{r_0:\overline{\omega - r_0|}} \tag{6.11.6}$$

Similarly with the argument of the stationary population and using the last equation (6.11.6) (which does not depend on the time variable t) we may distiguish two cases:

1st Case: F=0

Again following the reasoning of the stationary population and since equation (6.11.6) does not depend on the time variable the optimal path is defined from equations

$$c(t)=c_0 \quad \text{and} \quad r(t)=r_0 \quad \forall t \geq 0$$

and consequently $F(t)=0 \quad \forall t > 0$.

2^{nd} Case: $F \neq 0$

Under this case we have also two possible scenarios.

(i) The fund value F_0 is a small one (so similarly with the stationary population) we obtain again the optimal pair as for the 1^{st} case leaving the fund value constant i.e.

$$F(t) = F_0 \qquad \forall t > 0 \,,$$

(ii) The fund value of F_0 is <u>not</u> a small one. Again similarly with the respective case of the stationary population we choose initial values c(0) and r(0) slightly different from c_0 and r_0 in order to modify the fund value (reduce or increase it up or down to zero) while as the time passes (and as the fund goes to zero i.e. $F(t) \xrightarrow[t \to \infty]{} 0$) our choice will converge to c_0 and r_0 i.e.

$$\lim_{t \to \infty} c(t) = c_0 \quad \text{and} \quad \lim_{t \to \infty} r(t) = r_0$$

Non-Stable Population

In a non-stable population we may distinguish two cases.

1^{st} Case

The demographic pattern may be approximated by a combination of stationary or stable populations. For example, we may have a population with a decreasing growth rate say g(t) where

$$g(0) = g_0 \text{ and } \lim_{t \to \infty} g(t) = g_\infty \tag{6.11.7}$$

Consequently will be asymptotically stable with a growth rate of g_∞.

In such cases we may provide similar reasoning as before and imagine the optimal path lying in the area described by the lines (we have seen in the stable population) of the initial g_0 and ultimate g_∞ growth rates.

2^{nd} Case

The demographic pattern can not be described with a combination of the standard ones. Then in such cases we can not obtain a first taste but only solve the system (6.10.6) and design the respective optimal path according to the whole analysis described in section (6.10) for our model.

Stochastic pattern for population

The demographic pattern is a stochastic one. Then we should modify the equations and conditions of section (6.10) using expectations and variances for the respective variables of the system. Under this arrangement the minimization criterion will have the following format:

$$\min_{c(t),r(t)} \int_0^T E\left\{\theta[100(c(t)-c_0)]^2+(1-\theta)(r(t)-r_0)^2\right\}dt$$

Similarly we shall use in equations (6.10.1) and (6.10.5) the quantities

EF(t): expected value of the fund at time t

EB(t,x): expected value of B(t,x) at time t

EW(t,x): expected value of W(t,x) at time t.

6.12 The use of linear functions for Wages/Salaries (W) and benefits (B)

As we have already seen, our proposed model requires the use of functional analysis. Actually, in its general form described by the system (6.10.6) there is a question of a functional minimization. Consequently, the solution of the problem in its general form becomes very difficult. In order to tackle this difficulty and obtain a further insight in the model we shall discuss the use of linear functions for wages / salaries (W) and benefits (B). We shall provide the obvious reason why we may be allowed to consider linear functions in our development of the solution without losing the full generality or insight of our model.

The obvious reason is the existence of the linearization procedure of any function.

We recall the definition of W(t,x) and B(t,x) from section (6.10) i.e.

$$W(t,x) = \int_{a}^{x} p\ell(t,y) \cdot s(t,y)\,dy \tag{6.12.1}$$

and

$$B(t,x) = \int_{x}^{\omega} p\ell(t,y) \cdot b(t,y)\,dy \tag{6.12.2}$$

Now considering the most general format for the population $p\ell(t,y)$, the salary function s(t,y) and the benefit function (b(t,y) we obtain by integration W(t,x) and B(t,x) which may also have any general format.

Instead of these, we may use their linear approximations which are obtained by the standard linearization procedures. Firstly we have to choose the equilibrium point say (t_0,r_0) then we obtain

$$W(t,x)=W(t_0,r_0)+\frac{\partial}{\partial t}W(t,x)\Big|_{t=t_0}(t-t_0)+\frac{\partial}{\partial x}W(t,x)\Big|_{x=r_0}(x-r_0) \tag{6.12.3}$$

$$B(t,x)=B(t_0,r_0)+\frac{\partial}{\partial t}B(t,x)\Big|_{t=t_0}(t-t_0)+\frac{\partial}{\partial x}B(t,x)\Big|_{x=r_0}(x-r_0)\,. \tag{6.12.4}$$

Finally, using the linearization procedure for a series of points we may approximate our general functions of W and B with a family of linear functions.

6.13 The continuous form of the model [C-C] and the respective general solution

The [C-C] version of the model is the most general form of the problem. In order to handle it, we have to use functional minimization techniques. Let's consider the system of equations (6.10.6) i.e.

$$\min_{c(t),r(t)}\int_0^T\left\{\theta[100(c(t)-c_0)]^2+(1-\theta)\cdot[r(t)-r_0]^2\right\}dt \tag{6.13.1}$$

$$F'(t)=\delta\cdot F(t)+c(t)W(t,r(t))-B(t,r(t)) \tag{6.13.2}$$

$$F(0)=F_0 \quad\text{and}\quad F(T)=0 \tag{6.13.3}$$

Firstly, we shall not discuss the general sufficiency conditions for optimality for the problem above (these may be found in Athans & Falb (1966)). We shall use a result (see Kamien & Schwartz (1981)) which states that at least one solution exist if the integrand of (6.13.1) and the right hand side of differential equation (6.13.2) are convex in (F,c,r). Now the solution of the problem may be determined using the Hamiltonian of the system i.e.

The Hamiltonian is defined as:

$$\mathscr{H}(t)=\theta[100(c(t)-c_0)]^2+(1-\theta)\cdot[r(t)-r_0]^2+$$

$$+p(t)\delta F(t)+p(t)c(t)W(t,r(t))-p(t)B(t,r(t)) \tag{6.13.4}$$

where p(t) is the relevant costate vector of the system.

The optimal c(t) and r(t) controls can be found as the solution of the following system

$$\frac{\partial\mathscr{H}}{\partial c}=0 \text{ and } \frac{\partial\mathscr{H}}{\partial r}=0 \tag{6.13.5}$$

(sufficiency conditions for the existence of the minimum will be discussed in the next section). Differentiating the Hamiltonian according to (6.13.5) and equating to zero we obtain

$$2\theta 100^2[c(t)-c_0]+p(t)W(t,r(t))=0 \tag{6.13.6}$$

$$2(1-\theta)[r(t)-r_0]+p(t)c(t)\cdot\frac{\partial}{\partial r(t)}\cdot W(t,r(t))-p(t)\frac{\partial}{\partial r(t)}\cdot B(t,r(t))=0 \tag{6.13.7}$$

As we observe equation (6.13.6) may be solved with respect to c(t) and then be substituted in (6.13.7) eliminating the c(t). But the new equation which is produced from (6.13.7) is a partial differential equation with respect to r(t) and consequently not easily solveable.

Overcoming the last difficulty, we may generally write the solution as

$$C(t)=G_1(t,c_0,r_0,\theta,p(t)) \tag{6.13.8}$$

$$r(t)=G_2\left(t,c_0,r_0,\theta,p(t)\right) \tag{6.13.9}$$

(where G_1,G_2 are the suitable functions).

Now p(t) is given by the following differential equation

$$p'=-\frac{\partial \mathscr{H}}{\partial F} \tag{6.13.10}$$

or equivalently differentiating the Hamiltonian with respect to F we obtain

$$p'(t)=-\delta p(t) \Leftrightarrow p(t)=p_0 e^{-\delta t}\ , \tag{6.13.11}$$

Substituting (6.13.11) into (6.13.8) and (6.13.9) and then the resulting equation into the initial (6.13.2) we obtain a differential equation with respect to F(t).

Now solving this equation with respect to F(t) given the conditions (6.13.3) for the boundary values of F(0) and F(T) we may determine the p_0 which appears in equation (6.13.11).

Finally, the optimal path for c(t) and r(t) are defined by function G_3,G_4 as

$$c(t)=G_3\left(t,c_0,r_0,\theta,\delta,p_0\right) \tag{6.13.12}$$

$$r(t)=G_4\left(t,c_0,r_0,\theta,\delta,p_0\right)\ , \tag{6.13.13}$$

6.14 Special case for the [C-C] model using linear functions for W and B

As we have seen in the last section the general solution of the [C-C] model is quite complex, so we shall simplify it, by approximating with linear functions the total wages / salaries (W) and total benefits (B). Referring to the discussion of section (6.12), let's assume the following forms,

$$W(t,r(t))=\lambda_1 t+\lambda_2 r(t)+\lambda_3 \tag{6.14.1}$$

$$B(t,r(t))=k_1 t+k_2 r(t)+k_3 \tag{6.14.2}$$

where $k_1, k_2, k_3, \lambda_1, \lambda_2, \lambda_3$ are constant coefficients

(e.g. $\lambda_2 = \frac{\partial}{\partial r(t)} W(t, r(t))\Big|_{r(t)=r_0}$ etc... Substituting (while differentiating if necessary) equation (6.14.1) and (6.14.2) to (6.13.6) and (6.13.7) we obtain

$$2\theta 100^2 \cdot [c(t) - c_0] + p(t) \cdot [\lambda_1 t + \lambda_2 r(t) + \lambda_3] = 0 \tag{6.14.3}$$

$$2(1-\theta) \cdot [r(t) - r_0] + \lambda_2 p(t) \cdot c(t) - k_2 p(t) = 0 \ , \tag{6.14.4}$$

We shall rearrange the terms of the equation above in order to obtain the system in the standard format and solve it by using the relevant determinants i.e.

$$2\theta 100^2 c(t) + \lambda_2 p(t) r(t) = 2\theta 100^2 c_0 - \lambda_1 t p(t) - \lambda_3 p(t) \tag{6.14.5}$$

$$\lambda_2 p(t) c(t) + 2(1-\theta) r(t) = 2(1-\theta)\, r_0 + k_2 p(t) \ , \tag{6.14.6}$$

The solution of the system is given of equation (6.14.10) & (6.14.11) and provided as:

$$c(t) = \frac{D_{C(t)}}{D} \quad \text{and} \quad r(t) = \frac{D_{r(t)}}{D} \tag{6.14.7}$$

where

$$D = \det \begin{bmatrix} 2\theta 100^2 & \lambda_2 p(t) \\ \lambda_2 p(t) & 2(1-\theta) \end{bmatrix} = 4\theta(1-\theta) 100^2 - [\lambda_2 p(t)]^2 \ , \tag{6.14.8}$$

In order to obtain a unique solution $D \neq 0$ i.e.

$$p(t) \neq \pm \frac{200}{\lambda_2} \sqrt{\theta(1-\theta)} \ , \tag{6.14.9}$$

$D_{C(t)}$ and $D_{r(t)}$ are the determinants which are produced by substituting the column of the relevant index to the determinant D keeping also in mind equation (6.13.11) for the costate variable p(t) and substituting into the solution of c(t) and r(t) we finally obtain

$$c(t) = \frac{[2(1-\theta)] \cdot \left[2\theta 100^2 c_0 - \lambda_1 t p_0 e^{-\delta t} - \lambda_3 p_0 e^{-\delta t}\right] - \left[\lambda_2 p_0 e^{-\delta t}\right] \left[2(1-\theta) r_0 + k_2 p_0 e^{-\delta t}\right]}{4\theta(1-\theta) 100^2 - \left[\lambda_2 p_0 e^{-\delta t}\right]^2} \tag{6.14.10}$$

$$r(t)=\frac{\left[2\theta 100^2\right]\cdot\left[2(1-\theta)\,r_0+k_2p_0e^{-\delta t}\right]-\left[\lambda_2p_0e^{-\delta t}\right]\left[2\theta 100^2c_0-\lambda_1tp_0e^{-\delta t}-\lambda_3p_0e^{-\delta t}\right]}{4\theta(1-\theta)\cdot 100^2-\left[\lambda_2p_0e^{-\delta t}\right]^2}, \quad (6.14.11)$$

Equations (6.14.10) and (6.14.11) should be combined with equation (6.13.2) and using conditions (6.13.3) obtain the value of p_0.

Now we shall discuss the sufficiency conditions for the existence of the minimum.

The following matrix A_1 should be positive definite

$$A_1=\begin{bmatrix}\dfrac{\partial^2\mathscr{H}}{\partial c^2} & \dfrac{\partial^2\mathscr{H}}{\partial r\partial c}\\ \dfrac{\partial^2\mathscr{H}}{\partial c\partial r} & \dfrac{\partial^2\mathscr{H}}{\partial r^2}\end{bmatrix}=\begin{bmatrix}2\theta 100^2 & \lambda_2p(t)\\ \lambda_2p(t) & 2(1-\theta)\end{bmatrix}$$

i.e. the relevant determinant should be positive definite which is interpreted with the two inequalities (a) and (b) below

(a): $2\theta\cdot 100^2>0$

(b): $\det(A_1)=4100^2\theta(1-\theta)-[\lambda_2p(t)]^2>0$

The last inequality holds if and only if

$$p(t)\in\left(-\frac{200}{\lambda_2}\sqrt{\theta(1-\theta)},+\frac{200}{\lambda_2}\sqrt{\theta(1-\theta)}\right) \quad (6.14.12)$$

Closing this section we shall provide some comments with respect to the solution of the system (see equations (6.14.10), (6.14.11)) and its behavior according to the change of each parameter. Firstly about the ultimate values of our control i.e.

$$\lim_{t\to\infty}c(t)=c_0 \text{ and } \lim_{t\to\infty}r(t)=r_0\ , \quad (6.14.13)$$

Equation (6.14.13) may be easily proved considering the solution of c(t) and r(t) and since

$$\lim_{t\to\infty}e^{-\delta t}=0 \quad (6.14.14)$$

Hence, the ultimate values for c(t) and r(t) converge to the equilibrium point (c_0,r_0). Now as regards the behavior of the solution with respect to the different parameters involved we may draw some conclusions observing (a) the position of the parameter (i.e. whether it is in the numerator or denominator of the fractions in equations (6.14.10) and (6.14.11)) (b) the sign (positive or negative) and the relevant coefficient.

So according to those rules we may state that:

(1) Generally speaking, the two control variables c(t), r(t) have the same expression in their denominator while in their numerators have four brackets. Diagrammatically in the following sense.

$$c(t): \quad [A_1]\cdot[A_2]-[A_3]\cdot[A_4] \tag{6.14.15}$$

$$r(t): \quad [A_5]\cdot[A_4]-[A_3]\cdot[A_2] \tag{6.14.16}$$

From the last two expressions and as $[A_3]$ appears both in c(t) and r(t) with the same sign and the terms $[A_2]$, $[A_4]$ have opposite signs we may identify that c(t), r(t) are developed in opposite directions (when c(t) increases the r(t) decreases and vice versa).

(2) c_0 is in the numerator of c(t) with positive sign, so as this parameter increases the same happens to the magnitude of the solution for c(t). c_0 also appears in the numerator of r(t) but with the negative sign consequently the magnitude of r(t) but with the negative sign consequently the magnitude of r(t) will decrease as c_0 increases.

(3) A similar situation (as with comment (2)) applies for the r_0 parameter. As it increases then r(t) increases while c(t) decreases.

(4) As regards the δ force of interest we may say that the higher is the faster the solution converges to its ultimate state of (c_0,r_0).

(5) Parameters λ_1,λ_3 appear in numerator of c(t) with negative sign so the higher values will result smaller values for c(t). The opposite result holds for the r(t) as λ_1,λ_3 appears in numerator but with positive sign. This situation may be proved by general reasoning as λ_1,λ_3 appears in the function of wages. So the bigger λ_3 (constant term in function) or the bigger λ_1 (the slope of function with respect to time) the less contribution required.

(6) Parameter k_2 (the slope of benefits with respect to age of normal retirement) appears on the numerator of c(t) with negative sign and on the numerator of r(t) with positive sign (similar results as before may be drawn).

(7) Parameter λ_2 (the slope of wages with respect to age of normal retirement) appears both in numerator and denominator of c(t), r(t) with the same negative sign, so its effect is the same for the control variables while the direction of the effect depends on the magnitude of the parameters in the expressions.

(8) Finally parameter θ appears both in numerator and denominator of c(t) and r(t) (also the 2^{nd} power appears). From its pattern we may conclude that c(t) increases and r(t) decreases as θ increases (Expected also by general reasoning as θ is the weight of c(t)).

6.15 An hybrid [C-D] model and the respective general solution

The general model in the continuous form [C-C] (apart from being very difficult) it is not so practicable as describes the contribution rates and age of normal retirement with the continuous functions c(t) and r(t) respectively.

Obviously it is very difficult in practice to change these variables "continuously" (i.e. every week, or every day, or every hour!...). In this section we shall degrade our problem approximating better the real world. We shall observe our process at discrete time n=1,2,... Now we need to find the equivalent symbols and expressions for this type of model. Firstly, we have to find the equivalent annual accumulation factor J=1+i which is derived as

$$J = 1 + i = e^{\delta} , \tag{6.15.1}$$

Then the other symbols are defined similarly as in section (6.10) keeping in mind the new discrete format of the process (substitute variable t with n) i.e.

F_n: Reserve fund at time n (at the end of the n-th year).

c_n: Contribution rate during the n-th year, (constant for whole year)

r_n: Retirement age in the n-th year (constant for the whole year)

pl(n,x) , s(n,x) , b(n,x) , B(n,x) , W(n,x)

are defined similar as in section (6.10) with the consideration of the discrete variable n. So for example

B(n,x) stands for the benefits paid during year n if the age of normal retirement is fixed at x.

Now the basic differential equation (6.10.1.) becomes difference equation i.e.

$$F_{n+1} = F_n J + c_n W(n, r_n) - B(n, r_n) , \tag{6.15.2}$$

and the functional index of minimization (substituting the integral with the summation operator and taking the differences between consecutive values of c_n and r_n in order to determine a smooth path) i.e. the (6.10.3) expression becomes

$$\min \sum_{n=1}^{m} \left\{ \theta [100 (c_n - c_{n-1})]^2 + (1-\theta) [r_n - r_{n-1}]^2 \right\} , \tag{6.15.3}$$

We should state that expression (6.15.3) is very similar with the one used by Benjamin (1989) in the context of funding a defined pension scheme (in that approach m=4). Finally, expressions (6.10.3) and (6.15.3) are special cases of the general expression used by Haberman & Sung (1994), where the smoothing procedure asumes target values for each year n.

τr_n : target normal retirement age at time n

τc_n : target normal contribution rate at time n

So under the notation above we have that

- (6.10.3) uses $\tau r_n = r_0$ and $\tau c_n = c_0$ while
- (6.15.3) uses $\tau r_n = r_{n-1}$ and $\tau c_n = c_{n-1}$

We should also mention that expression (6.15.3) differs from expression (6.10.3) in the sense that the former smooths the path for c_n and r_n using the first differences while the latter smooths the path for c(t) and r(t) using the differences of these values from some standard ones (i.e. c_0 and r_0).

The equivalent boundary conditions are the following

$$F_0 \quad \text{and} \quad F_m=0 \ , \tag{6.15.4}$$

In order to proceed with the solution of the model we shall combine the difference equation (6.15.2) with (6.15.4) and produce one "large" equation as below

$$F_0 J^m + \sum_{n=1}^{m} \left[c_n W(n,r_n) - B(n,r_n)\right] J^{m-n} = F_m = 0 \ , \tag{6.15.5}$$

Now observing the last equation and the expression (6.15.3) we conclude that our minimization problem requires lagrange multipliers (i.e. it is an optimization problem with constraints). Hence, the problem is restricted to the expression below:

$$\min \Lambda(c_1,\ldots,c_m,r_1,\ldots,r_m,\lambda,c_0,r_0,\theta,J,F_0) \ , \tag{6.15.6}$$

where Λ should be minimized with respect to $\lambda, c_1, c_2, \ldots, c_m, r_1, r_2, \ldots, r_m$ and is the following function

$$\Lambda = \sum_{n=1}^{m} \left\{ \theta[100(c_n - c_{n-1})]^2 + (1-\theta)[r_n - r_{n-1}]^2 \right\} +$$

$$+\lambda \left\{ F_0 J^m + \sum_{n=1}^{m} [c_n W(n, r_n) - B(n, r_n)] J^{n-m} \right\} \tag{6.15.7}$$

So the function is a real-valued function with (2m+1) free variables. Its minimization requires partial differentiation to each of the (2m+1) variables then equating to zero and finally solve of the system of (2m+1) simultaneous equations i.e.

$$\frac{\partial \Lambda}{\partial c_i} = 0 \text{ and } \frac{\partial \Lambda}{\partial r_i} = 0 \text{ and } \frac{\partial \Lambda}{\partial \lambda} = 0 \qquad \text{where } i=1,2,\ldots,m \tag{6.15.8}$$

As we observe the general solution is again quite complex, so we shall consider in the next section linear functions for W and B in order to obtain an analytical expression for the partial derivatives of the system (6.15.8).

6.16 Special case for the [C-D] model using linear functions for W and B

Similarly with section (6.14) we assume the linear functions for W and B given by equation (6.14.1) and (6.14.2) (in the discrete form) i.e.

$$W(n, r_n) = \lambda_4 n + \lambda_5 r_n + \lambda_6 \tag{6.16.1}$$

$$B(n, r_n) = k_4 n + k_5 r_n + k_6 \tag{6.16.2}$$

Now according to the system (6.15.8) we obtain,

$$\left.\begin{array}{ll}
200\theta(c_i - c_{i-1}) - 200\theta(c_{i+1} - c_i) + \lambda[\lambda_1 i + \lambda_2 r_i + \lambda_3] = 0 & \\
2(1-\theta)(r_i - r_{i-1}) - 2(1-\theta)(r_{i+1} - r_i) + \lambda[\lambda_2 c_i - k_2] = 0 & \\
F_0 J^m + \sum_{n=1}^{m} [c_n W(n, r_n) - B(n, r_n)] J^{n-m} = 0 & i=1,2,\ldots,m
\end{array}\right\} \tag{6.16.3}$$

Unfortunately we arrive again at (6.11.3) a quite complex system of simultaneous equation which may be solved only by numerical methods.

6.17 The discrete form of the model, called [D-D] and the respective algorithm

The two previous models [C-C] and [C-D] are interesting from the theoretical point of view but may not be applicable in practice, as require the continuous form for the functions of wages W and benefits B, over the interval [O,T].

In this section we shall develop another version of the basic model, called [D-D] which may easily applicable to real data. We shall consider that functions of W and B are defined (by some kind of projection) for discrete values of (n) and (x). So actually we have a table with respect to each function.

Table (6.17.1)

x \ n	0	1	2	...
65	...		...	
66	...		...	
67	Projected values of W(n,x)			
⋮	...		...	

Table (6.17.2)

x \ n	0	1	2	...
65	...		...	
66	...		...	
67	Projected values of B(n,x)			
⋮	...		...	

Similarly with section (6.15) equation (6.15.2) also holds in this version of the model i.e.

$$F_{n+1} = F_n(1+i) + c_n W(n,r_n) - B(n,r_n) \qquad (6.17.1)$$

Now in order to find the optimal ("smoothest") path we shall simulate different paths for (c_n, r_n) under the following simple rules

$$c_{n+1} = \begin{cases} c_n + \Delta c \\ c_n \\ c_n - \Delta c \end{cases} \qquad (6.17.2), \qquad r_{n+1} = \begin{cases} r_n + \Delta r \\ r_n \\ r_n - \Delta r \end{cases} \qquad (6.17.3)$$

i.e. consider small changes Δc and Δr starting from (c_0,r_0) and finally after m years calculate the value of the functional of expression (6.15.3) i.e.

$$\sum_{n=1}^{m} \left\{ \theta[100(c_n - c_{n-1})]^2 + (1-\theta)\,[r_n - r_{n-1}]^2 \right\} \qquad (6.17.4)$$

Finally, we shall relax the second condition of (6.15.4) (which states that the final value F_m should be zero) and demand only that F_m should be "near" to zero i.e.

$$F_m \in (-\gamma,+\gamma) \qquad (6.17.5)$$

(where γ is a small positive number),

Then we shall collect all the paths which produce a final value F_m in the required interval $(-\gamma,+\gamma)$ and examine which minimizes expression (6.17.4).

Hence, that is going to be the optimal path. Of course the whole procedure is not simple! Assuming that we want to find the optimal path for the next m years (i.e. determine the path from (c_0,r_0) up to (c_m,r_m)) then we should simulate and check 9^m paths!!! (The proof for the number of the paths is obvious because having defined (c_n,r_n) the next possible positions for (c_{n+1},r_{n+1}) are actually nine. Using standard theory of combinations we obtain the last number of 9^m paths.)

Now it is more clear that the solution by the simulation techniques is not just "simple" but for large values of m (e.g. m=20 or 30 which may be practicable) the solution procedure is quite heavy even for the fastest computer.

For the reason above and in order to obtain a solution for the real data used in the next sections we shall modify this algorithm by drawing some additional comments about the demographic pattern and for the development of the reserve fund.

Closing this section, we shall provide a further notice for the small changes Δc and Δr which appear in the algorithm for the simulation of the path for c_n and r_n.

These quantities should be small (depending also on the selected time unit, if we assume that the time unit is a calendar year then small means

$$0<\Delta c\leq 1\% \tag{6.17.6}$$

and

$$0<\Delta r\leq 1 \tag{6.17.7}$$

(in the last inequality 1 stands for one year of age).

We must also notice that some relationship should exist between the choices for Δc and Δr. A full investigation of the problem may require several simulations with different values for the pair of Δc and Δr in order to decide which is the optimal path for the contribution rate and age of normal retirement. For example the first simulation may consider annual changes of 0.25% for the contribution rate and 0.25 (i.e. three months) for the age of normal retirement. The second simulation may consider annual changes of 0.5% and 0.5 (i.e. six months) respectively and so on. Finally we may be able to choose the optimal path of all the set of simulation.

6.18 General description of the demographic pattern in Greece (Projections up to 2020)

We have already seen in section (6.2) that Greece follows the international demographic trend of the "aging populations". As we observe from table (6.2.2) the elderly dependency ratio for those lives aged 65 and over will steadily be increased the next 20 years (from 15.3% in 1995 will go up to 20.0 in the year 2015).

It has also the standard characteristics of an "aging population" i.e. a decreasing pattern for fertility rates while increasing for life expectancy. A further insight may be gained by considering table (6.18.1) which presents a projection (using statistical data up to 1995) up to the year 2020. It will be also helpful to observe the 3-D diagram (6.18.1) which presents the data of table (6.18.1) from the age band [25-29] up to [90+...]. Considering the normal practice where a young person enters the labour force approximately at age 25 (after completing his studies and the compulsory service of 1.5 years to the Greek Army etc.).

Then diagram (6.18.1) presents both the active group and the group of retired lives. As we can see the graphs of the first four age bands i.e. [25-29], [30-34], [35-39], [40-44] generally decreases while all the remaining graphs of the respective age bands have an increasing trend. We may also observe from table (6.18.1.) that the total population almost remains constant (actually has a marginal increase) near to 10.5-11 millions.

The Greek government has attempted (in the last 15 years) to provide some incentives (financial) in order to increase the fertility rates but they were unsuccessful up to now. Consequently, under the situation described above and the even worse perspectives with respect to the funding process of the Social Security System, government has started a dialogue process with all the parties involved in the funding process of the system.

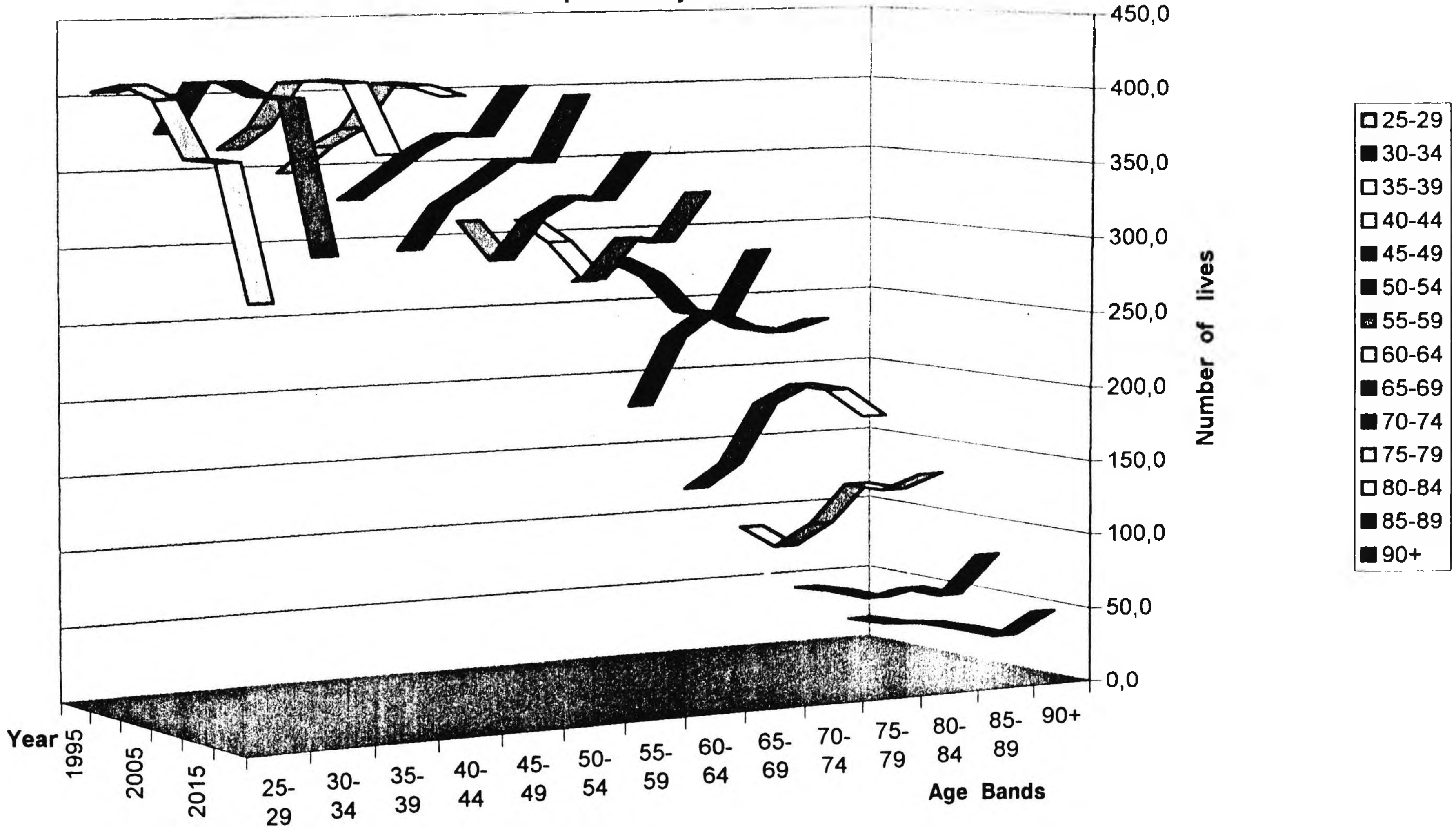
Although the system has already been reformed once in 1991 the proposed changes appear now non-adequate in order to solve efficiently the funding problem at the beginning of the next century.

Table (6.18.1)

AGE BANDS	1995 MALES	1995 FEMALES	2000 MALES	2000 FEMALES	2005 MALES	2005 FEMALES	2010 MALES	2010 FEMALES	2015 MALES	2015 FEMALES	2020 MALES	2020 FEMALES
0-4	268,9	253,1	285,3	267,9	300,9	282,5	298,1	279,9	289,9	272,2	285,1	267,6
5-9	294,1	278,3	271,6	256,0	286,5	269,3	302,1	283,9	299,4	281,3	291,2	273,6
10-14	357,7	337,8	298,1	282,0	273,5	257,8	288,4	271,0	304,0	285,7	301,3	283,1
15-19	395,3	373,5	361,6	341,4	300,0	283,8	275,4	259,6	290,3	272,9	305,9	287,5
20-24	405,0	385,6	399,9	379,4	363,3	344,2	302,0	286,7	277,7	262,6	292,5	275,8
25-29	403,4	394,7	409,8	391,1	401,6	381,9	365,2	346,7	304,3	289,4	280,1	265,3
30-34	375,3	376,1	410,1	400,9	412,4	393,9	404,2	384,6	368,1	349,6	307,4	292,4
35-39	363,4	364,7	379,8	380,1	411,3	402,3	413,5	395,4	405,6	386,2	369,7	351,3
40-44	346,6	343,1	365,9	367,5	379,7	380,9	411,0	403,1	413,5	396,3	405,6	387,2
45-49	327,6	327,7	345,0	342,5	363,2	366,2	376,9	379,5	408,3	401,8	410,8	395,0
50-54	292,0	300,1	324,9	327,5	341,2	341,2	359,0	364,7	373,2	378,2	404,0	400,3
55-59	310,6	330,0	286,3	298,6	317,6	324,7	333,5	338,3	351,9	361,9	365,6	375,3
60-64	309,8	332,2	297,1	324,3	274,7	293,5	304,6	319,1	321,2	333,0	338,9	356,2
65-69	265,8	301,3	283,6	318,1	275,2	312,3	254,6	282,7	284,3	308,4	299,9	321,9
70-74	178,1	222,1	228,7	275,6	249,4	294,9	241,9	289,4	226,9	264,0	253,3	288,1
75-79	117,7	157,0	139,5	188,0	185,0	238,4	201,5	254,9	199,7	253,7	187,5	231,7
80-84	87,4	123,4	78,5	112,7	99,0	142,1	131,1	179,9	147,7	197,7	146,2	196,5
85-89	42,3	57,6	44,6	67,2	44,7	67,8	56,5	85,6	79,2	114,0	89,0	125,0
90+	17,0	26,2	18,9	28,2	24,1	38,1	25,5	41,2	34,1	54,9	47,4	73,1
TOTAL	5.158	5.285	5.229	5.349	5.303	5.416	5.345	5.446	5.379	5.464	5.381	5.447
GRAND T	10.443		10.578		10.719		10.791		10.843		10.828	

Source : National Statistical Service of Greece (1995)

Diagram (6.18.1)
A projection for the Greek population
up to the year 2020



Now the question is: which is the solution? We may argue as before and finally propose the model of this chapter as a small contribution to the large problem of Social Security System.

Of course, our simulation will still be theoretical without taking in account all the complexities involved. We shall use table (6.18.1) assuming that we stand on a equilibrium point and then we shall design a potential path for c_n and r_n over the next years up to the year 2020.

Finally as regards the reliability of the data of table (6.18.1) we must stress that the projection has been based on the most recent experience (up to 1995) of the whole population of Greece. So we may assume full credibility to the figures shown in the mentioned table.

6.19 Application of the [D-D] model to the projected population of Greece and the respective simulation algorithm

In this section, we shall describe the respective algorithm (data, assumptions, equations, techniques etc.) used for the simulation of the [D-D] model to the projected population of Greece shown in table (6.18.1). The algorithm may be described by the following ten model assumptions:

1st Assumption

Firstly, we identify the potential trend of the optimal path for c_n and r_n. As we have comment in section (6.18) Greece exhibits the international demographic trend of "aging populations" (the elderly dependency ratio is shifting to the right without any fluctuations to the left side of the age band). So actually the potential optimal path should have a steadily increasing trend. Under this situation we may modify equation (6.17.2) and (6.17.3) reflecting the specific trend

$$c_{n+1} = \begin{cases} c_n + \Delta c \\ c_n \end{cases} \qquad (6.19.1) \qquad\qquad r_{n+1} = \begin{cases} r_n + \Delta r \\ r_n \end{cases} \qquad (6.19.2)$$

As a first step this modification will degrade the number of potential simulated paths down to 4^m (in section (6.17) were 9^m paths).

2^nd^ Assumption

The starting point of our simulation is the 1st of January 1998 (i.e. the first year (n=1) is the calendar year 1998) and the ending point is the 31st of December 2020 (i.e. the m-th last year is the calendar year 2020), consequently m equals 23 (m=23). That means we must simulate 4^{23} paths, (where 4^{23} equals to some hundreds of trillions!!!). But we are not going to do so because,

a) that was going to be a very heavy procedure even for a powerful computer station.

b) (more important) we have to consider some kind of boundaries for the values of the reserve fund i.e. if we leave the length of simulation equal to m=23 then the reserve fund will fluctuate deliberately (as there is no restriction) for a very long time up to the ending point where we should have put a condition similar to condition (6.17.5).

Considering the two reasons above we split the simulation period into four subperiods (i.e. 6,6,6,5 years respectively) applying the modified [D-D] algorithm to these subperiods obtaining the optimal path in four pieces. Analytically the four subperiods contain the following calendar years.

1st subperiod :		1998 - 2003	inclusively
2nd	" " :	2004 - 2009	" "
3rd	" " :	2010 - 2015	" "
4th	" " :	2016 - 2020	" "

At the end of each subperiod there is a condition for the reserve fund similar to condition (6.17.5) i.e.

$$F_6 \in (-\gamma_1,+\gamma_1),\ F_{12} \in (-\gamma_2,+\gamma_2),\ F_{18} \in (-\gamma_3,+\gamma_3),\ F_{23} \in (-\gamma_4,+\gamma_4)$$

where $\gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \gamma_4$, (6.19.3)

So actually with the existence of the last condition (6.19.3), the reserve fund is forced steadily and "smoothly" (depending on the choice for the values of $\gamma_1,\gamma_2,\gamma_3,\gamma_4$) to zero. In our application the choice for γ_i's i=1,2,3,4 is

$$\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 10 \tag{6.19.4}$$

The units of γ_i's are compatible with the monetary units of W(t,x) and B(t,x) (see the 5th Assumption). Finally we should stress that the certain choice for γ_i's (equal to 10) was decided in order to obtain a solveable system i.e. For smaller values (less than 10) condition (6.17.5) was not ultimately reachable for all the four subperiods.

3rd Assumption

The required tables (similar to tables (6.17.1) & (6.17.2)) for functions W(n,x) and B(n,x) are obtained from table (6.18.1) using linear interpolation in order to obtain the values for the intermediate values of the calendar years (e.g. 1998, 1999, 2001, 2002,... etc.) and for the whole age band (integer of fractional ages e.g. 66.5, 66.0, 66.5, 66,0, ... etc.).

4^{th} Assumption

Active lives are considered all lives (males and females) aged 25 and above up to the relevant retirement age limit. (Actually the implicit assumption for the labour force participation rate equals to 100%).

5^{th} Assumption

Assume that all active lives receive an annual income (wage / salary) equal to one money unit (on which the contribution rate is applicable) and retirees receive a pension of one half of money unit. This assumption may comply with the typical situation in Greece for the last decade where a pensioner receive a pension equal to 50% of the current salary of an active life.

6^{th} Assumption

The initial conditions for the first subperiod are:

$$c_0=15\%,\quad r_0=65,\quad F_0=0 \tag{6.19.5}$$

The conditions for the contribution rate and the age of normal retirement have been chosen such that their application to the calendar year 1997 will result a zero balance for the basic equation (6.3.1) in that year. The initial conditions for the 2nd, 3rd and 4th subperiods are determined consecutively from the determination of the optimal path of the previous subperiods (1st, 2nd, 3rd respectively).

7^{th} Assumption

The increases Δc and Δr have been fixed to the following figures:

$$\Delta c=0.5\% \quad (6.19.6), \qquad \Delta r=0.5 \text{ (six months)} \quad (6.19.7)$$

The values above may be considered adequately small for a "smooth" path.

8^{th} *Assumption*

The weights for c_n and r_n are determined by θ which is fixed to 0.50

$$\theta=0.50 \tag{6.19.8}$$

The conditions above may be interpreted: People are indifferent to a Δc increase in the contribution rate or a Δr increase in the age of eligibility (retirement age).

9^{th} *Assumption*

The interest rate is chosen to be a small one

$$i=4\% \tag{6.19.9}$$

10^{th} *Assumption*

The design of the optimal path for the 1st, 2nd and 3rd subperiod requires the simulation of $4^6=4096$ potential paths for (c_n,r_n) while the 4th subperiod requires $4^5=1024$ paths. Equation (6.17.1) is used to obtain the values for the reserve fund while expression (6.17.4) is minimized under the conditions described in 2nd step for the values $(F_6,F_{12},F_{18},F_{23})$ of the reserve fund.

The whole program is developed under the platform of a spreadsheet (Microsoft Excel).

The results of our simulations are provided into the following pages in table (6.19.1) and diagrams (6.19.1), (6.19.2), (6.19.3), (6.19.4) and (6.19.5). Full explanation and critical comments for the table & diagrams will be provided in the next section (6.20).

Table (6.19.1)
(Optimal Path)

Year	D.R. (65+)	c_n	r_n	F_n
1998	23,41%	15,0%	65,0	-15,80
1999	23,70%	15,0%	65,5	-7,71
2000	23,99%	15,5%	65,5	15,60
2001	24,28%	15,5%	65,5	25,00
2002	24,57%	15,5%	65,5	20,10
2003	24,86%	15,5%	65,5	0,26
2004	25,13%	16,0%	65,5	-6,04
2005	25,41%	16,0%	65,5	-27,20
2006	25,48%	16,0%	65,5	-53,60
2007	25,54%	16,5%	65,5	-56,10
2008	25,61%	16,5%	66,0	-25,90
2009	25,67%	16,5%	66,0	0,47
2010	25,74%	16,5%	66,0	22,90
2011	26,00%	16,5%	66,0	33,90
2012	26,26%	16,5%	66,0	33,00
2013	26,52%	16,5%	66,0	19,60
2014	26,78%	16,5%	66,0	-6,72
2015	27,03%	16,5%	66,5	-7,03
2016	27,30%	16,5%	66,5	-20,02
2017	27,57%	16,5%	66,5	-46,80
2018	27,83%	16,5%	67,0	-46,80
2019	28,10%	16,5%	67,5	-18,30
2020	28,37%	16,5%	67,5	-0,71

Diagram (6.19.1)

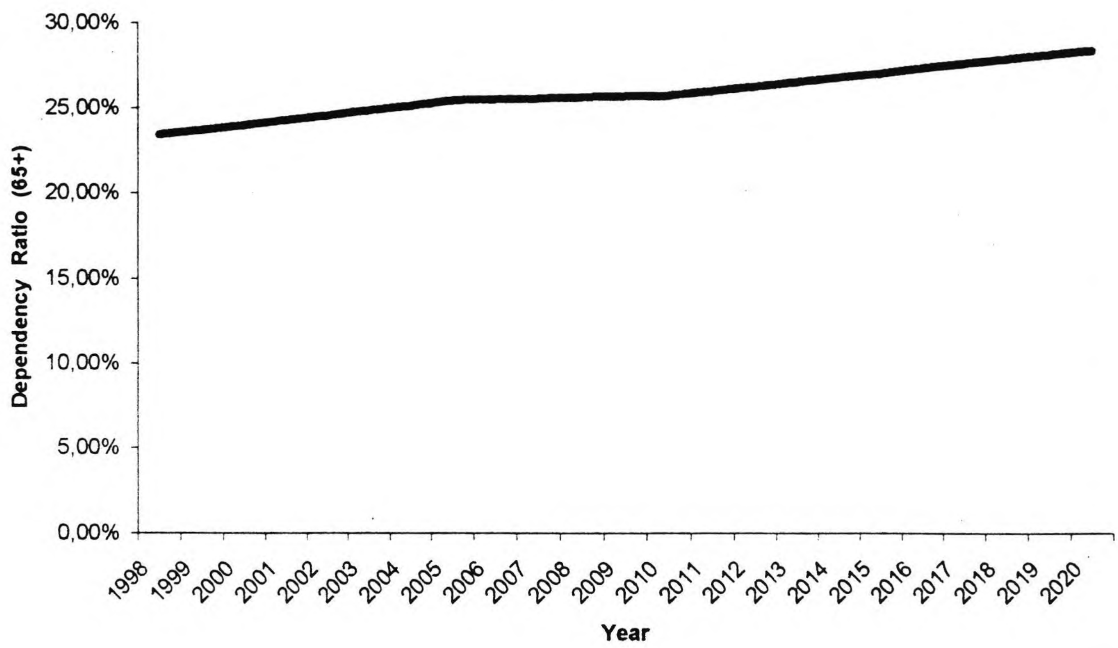
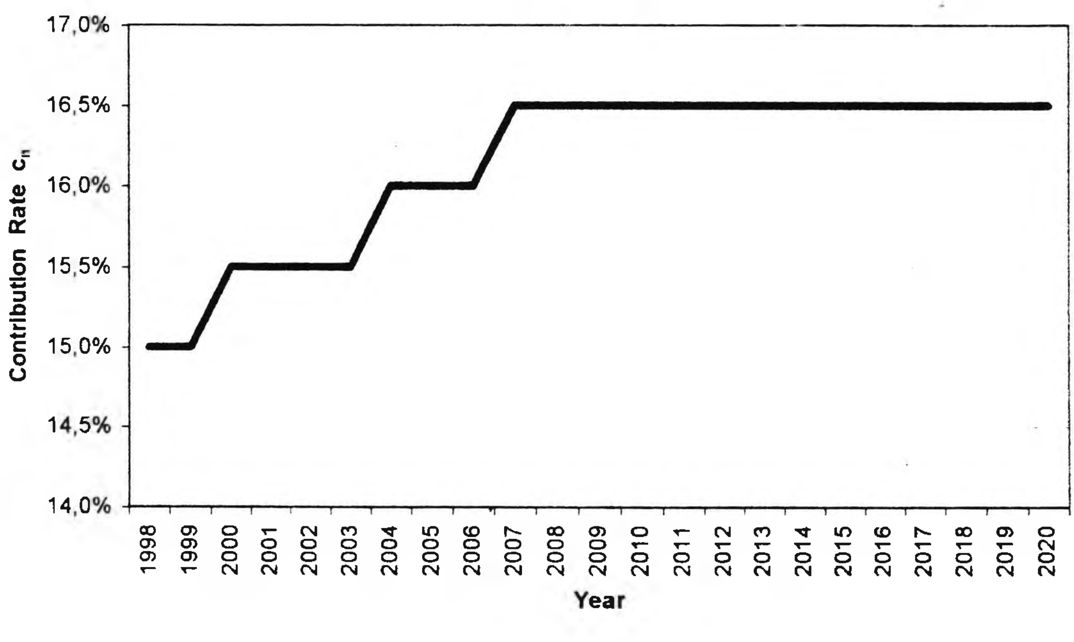


Diagram (6.19.2)



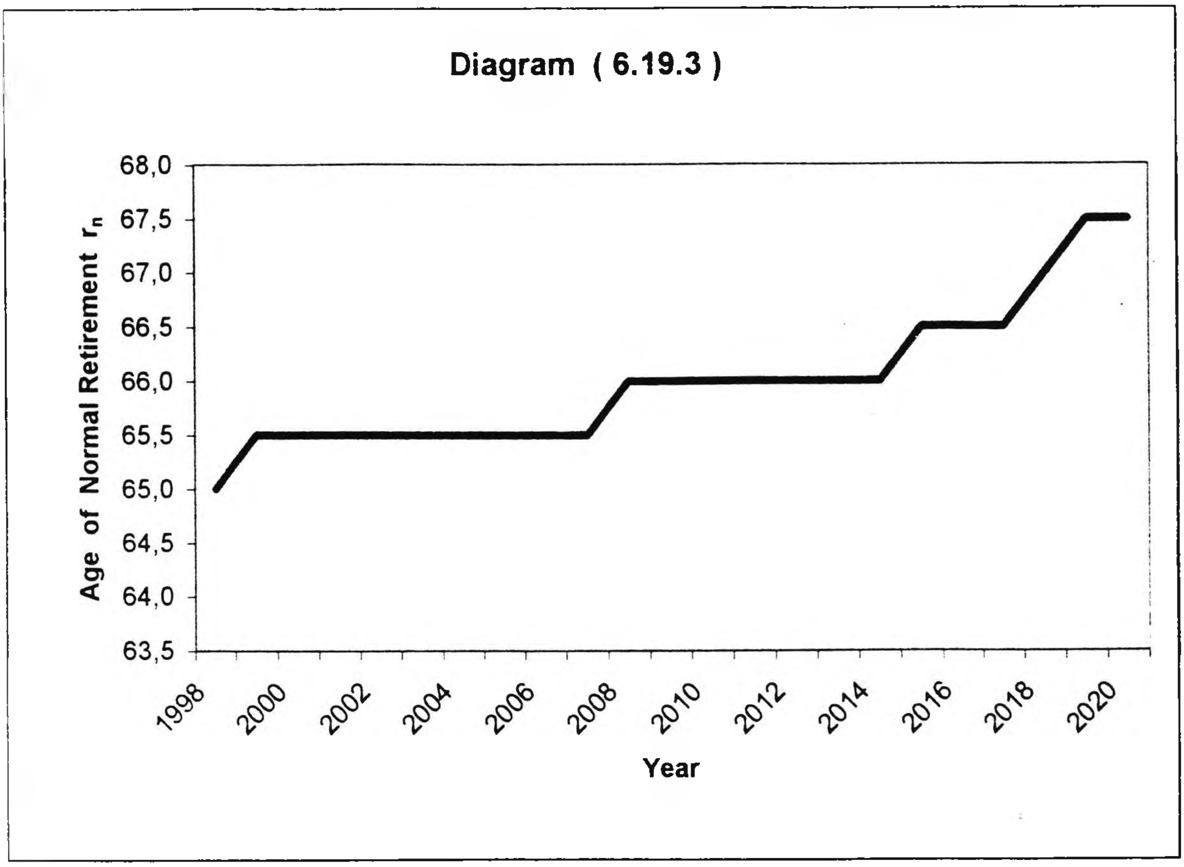


Diagram (6.19.3)
68,0
67,5
67,0
66,5
66,0
65,5
65,0
64,5
64,0
63,5
Age of Normal Retirement r_n
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

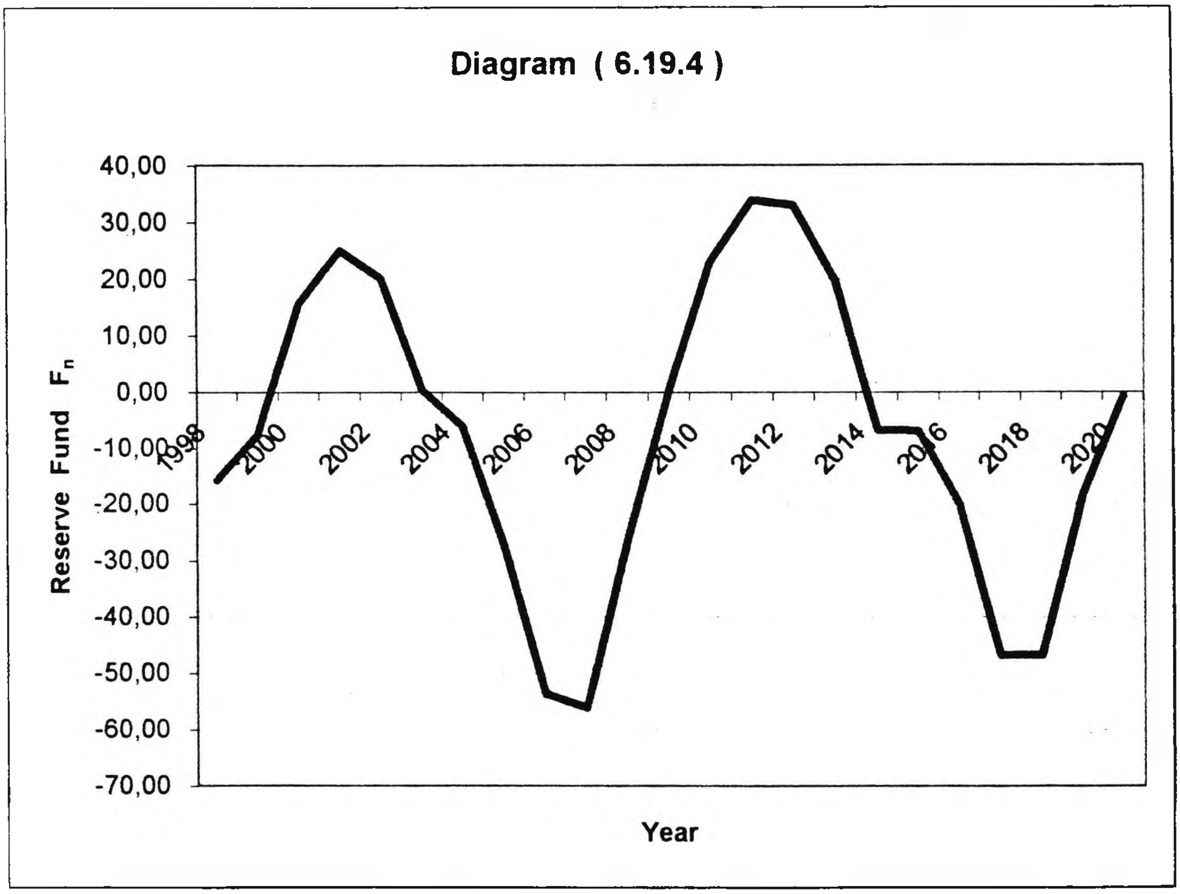
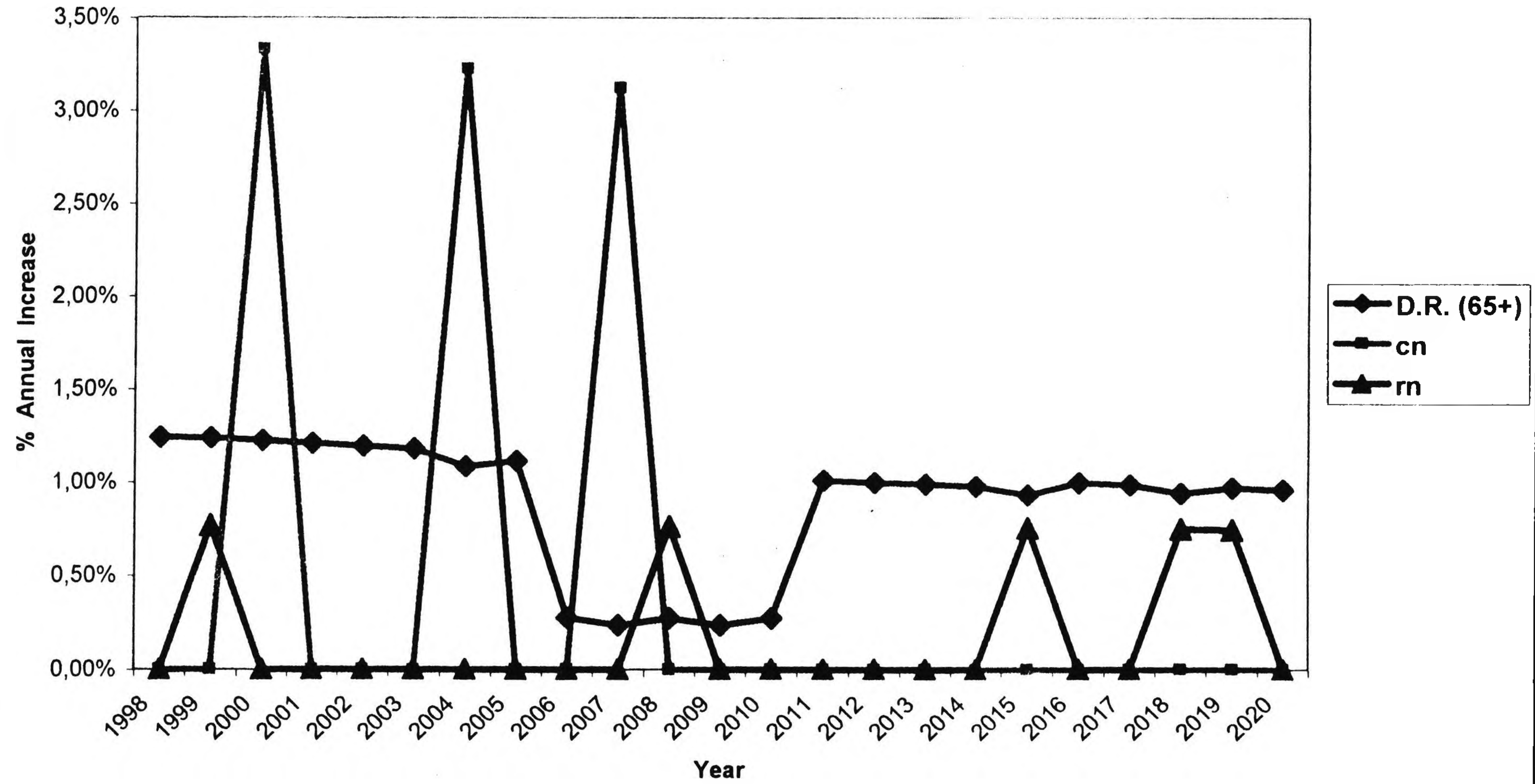


Diagram (6.19.4)
40,00
30,00
20,00
10,00
0,00
-10,00
-20,00
-30,00
-40,00
-50,00
-60,00
-70,00
Reserve Fund F_n
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

Diagram (6.19.5)



6.20 Comments on the results of our simulations for the actual application of [D-D] model

In this section, we shall comment on the structure and the results from our simulations which were presented in table (6.19.1) and diagrams (6.19.1), (6.19.2), (6.19.3), (6.19.4) and (6.19.5).

Table (6.19.1) contains the values of the optimal path for c_n and r_n contribution rate & normal retirement age for the first year which is actually the calendar year 1998 and up to 23rd year i.e. the calendar year 2020 (see the 3rd and the 4th column). It also presents the relevant values for the reserve fund at the end of each year (see the 5th column, F_n) and the development of the elderly dependency ratio D.R. (65+) (see 2nd column) which is defined as:

$$\text{D.R.}(65+) = \frac{\text{Total lifes aged 65 and over}}{\text{Total lives aged 25 and over}} \tag{6.20.1}$$

This is a slightly different definition than the normal one where the denominator equals to "total lives aged 25 up to 65".

(We consider lives 25 and over in order to focus only in the development of the labour force volume and not in the development of the young ages below 25).

Diagrams (6.19.1), (6.19.2), (6.19.3) and (6.19.4) present the development of the 2nd, 3rd, 4th and 5th column of the table (6.19.1) respectively.

Diagram (6.19.5) presents the annual increase (as a percentage of the previous value) for the 2nd, 3rd and 4th column of table (6.19.1). Actually the quantities presented are given by the formulae

$$\frac{\text{D.R.}(65+)_{n+1} - \text{D.R.}(65+)_n}{\text{D.R.}(65+)_n}, \quad \frac{c_{n+1} - c_n}{c_n}, \quad \frac{r_{n+1} - r_n}{r_n} \tag{6.20.2}$$

and may be considered as the graph of the relevant slope for Diagrams (6.19.1), (6.19.2) and (6.19.3) respectively.

Now as regards the form of the table and diagrams we state the following observations.

1) From the 2nd column of table (6.19.1) and the relevant diagram (6.19.1) we observe a steady increase (marginal) for the DR(65+). That was expected as this is the basic characteristic of an "aging population". And we can also see from table (6.19.5) the slope of this increase is almost constant for the whole period except a break in the years 2006-2010 where is near zero. (That means for those five years the phenomenon of aging populations slows down its speed.)
2) From the 3rd column of table (6.19.1) and the relevant diagram (6.19.2) we observe that c_n is steadily increased for the first half of the period 1998-2008 from 15% up to 16.5% and then remains constant to that value for the remainder of the period up to the year 2020. The opposite pattern appears for the development of r_n (see 4th column of table (6.19.1) and the relevant diagram (6.19.3)) i.e. constant pattern (or marginal increase) for the first half period (1998-2008) and increasing there after (rapidly after the year 2015).

   That was expected as the two variable c_n and r_n compete with each other. So for the first half of the period where we could say that the phenomenon of "aging population" is not so heavy, the steady increase in the contribution rate is adequate to balance the system without shifting the age of eligibility. In a small period between the years 2008-2015 we observe that both variables remain constant and this may be attributable to the lower speed in the increase of D.R. (65+) (see Diagram (6.19.5) for the relevant years). From the year 2015 up to the end the phenomenon of the aging population is so heavy that the system achieves its balance by shifting rapidly the normal retirement age.

3) From the 5^{th} column of table (6.19.1) and the relevant diagram (6.19.4), we observe that the reserve fund fluctuates rapidly with a period almost the half of our investigation period i.e. the period should be equal to 10-14 years. That may reveal a relevant pattern for the basic characteristics of the population i.e. the fertility rates and the life expectancy which appears to have a cyclical pattern every 5-7 years.
4) Finally from diagram (6.19.5) we may observe that there is a relationship between the patterns of the slopes. Generally we can say that the graphs for the slope of contribution rate and normal retirement age react with a delay (of almost 3 years) to the relevant pattern of the dependency ratio. This is quite clear in the middle of the investigation period (see the years 2006-2010 for the dependency ratio and 2009-2014 for the contribution rate & normal retirement age.) Additionally, we may also observe that the area described from the graph of the dependency ratio and the x-axis equals (approximately) the sum of the two areas described from the graph of the contribution rate with x-axis and the graph of the retirement age and x-axis. That may also be argued by general reasoning saying that the increase in contribution rates and normal retirement age should fully absorb the increase in the dependency ratio (which is actually a combination of the decline in fertility rates and increase in life expectancy).

6.21 Conclusions

In this chapter, we have attempted to obtain a deeper insight into the mechanisms and underlying philosophy of the PAYG model. Having identified the international demographic trend of "aging populations" and the respective problems (which may lead to a big crisis), we discussed different solutions described under a control theoretical context.

Basically, we have examined the three major variables of the PAYG model, i.e. the reserve fund, the contribution rate and the age of normal retirement.

Firstly, we show that "inter-generational equity" concept may be well served under the existence of a contingency fund (see section (6.9)). The fund absorbs the random fluctuations in the mortality pattern or the fertility rates and consequently smooths in a better way the contribution rates and the rates of return for each cohort of lives.

Secondly, we have investigated the different links of the contribution rate with the fertility rates and age of normal retirement with the life expectancy in order to realize the "inter-generational equity" concept. It is also has been pointed that people's expectations should be investigated and be incorporated (adjusting the weights of the two control variables mentioned before) in the objective function (section (6.9)).

Under the lines described above we have constructed a general model ([C-C] version) in order to design an optimal control path for the contribution rate and age of normal retirement. We have overcome the difficulty of the complicated version of the problem by considering linear functions for wages/salaries and benefits obtaining an analytical solution for the two control variables.

Then the initial model is simplified more ([C-D] version) obtaining a similar approach to Benjamin (1989). The full solution of this version required numerical methods.

Finally, we have introduced an additional simplification to our model ([D-D] version) trying to obtain a realistic algorithm which may be used in practice, especially for some statistical data of the Greek population.

The application of the third version of the model, in the Greek population may be considered successful. We have obtained the optimal path for contribution rates and retirement age for a projection period of 23 years. Starting in 1998 with a pair of values equal to (15%, 65) (for c_n and r_n) we have ended in 2020 with a pair of (16.5% 67.5). Insight has also been obtained into the relationship between the variables and their trends using the relevant diagrams of section (6.19).

Chapter 7

Conclusions

7.1 Finally "everything is under control"! (Stability instead of ruin)

At this stage of this thesis, we have put "everything under control". We have examined three insurance problems:

(a) a non-life insurance portfolio

(b) a special reinsurance arrangement (multinational pooling)

(c) the PAYG funding method

and have proved the powerfulness of the new approach using control theory.

Our basic effort for every problem was to design a stable system and thereafter to find the optimal control or the optimal choice for the parameters involved.

In the control approach the critical concept of "ruin" is replaced with the stability concept. We are not "afraid of ruin" (as described in the traditional approach) and so no attention is paid with respect to its calculation. But we are "afraid of instability" (i.e. the situation where the system diverges to infinity) and so, careful design is employed in order to achieve the desired level of stability.

Actually we should redefine the ruin (in the control theory context) as instability.

7.2 A short review of all the important results of the thesis

In this section, we shall try to outline the most important results of the thesis. The presentation will follow the order of the chapters. Hence,

Chapter 2

It is quite clear that control theory concepts may well match with the needs of the actuarial science. The successful application of control theory to a specific insurance problem encourages for further analysis & research.

Chapter 3

The research up to now has been restricted mostly in conventional control theory with some exceptions focused in modern optimal control techniques. Having identified this point, the concern of the current thesis was employed in modern control theory (i.e. examining time-varying problems & multiple input - multiple output models).

Chapter 4

The most important results in Chapter 4 which discusses the general problem of insurance pricing are:

(a) The existence of a critical value (f_∞) for the delay factor where beyond that value the system becomes unstable irrespective of the choice of the (ε) profit sharing feedback factor. That is the critical point where the information of the past becomes completely useless. The value of f_∞ equals to the perpetuity (in arrears) value at a rate of interest j=R-1 i.e.

$$f_\infty = \frac{1}{R-1} = a_{\overline{\infty|}}$$

(b) As regards the simulation results, it was very interesting to identify the potential minimum for the variance of the output response, with respect to the (ε) feedback factor in the area of 50% - 60%.

(c) Finally, in the time-varying model we have fully examined the stability condition in conjunction with the boundaries of Ω. (Ω is the set of possible values for R_n, n=1,2,...) i.e. The value of the upper bound of Ω determines the choice of (ε) feedback factor in order to produce a stable system.

Chapter 5

In chapter 5, we have consider a special reinsurance arrangement (multinational pooling) under three versions of a basic model. The important results are the following.

(a) The potential points for stability analysis (i.e. equilibrium points) are restricted to zero value given that $\varepsilon_1 \neq 0, \varepsilon_2 \neq 0, \ldots, \varepsilon_m \neq 0$ (non-negative feedback for each system).

(b) All the three versions are complete controllable so optimal solutions may be designed. Optimal values have been discussed for the parameters of the models either with analytical or numerical methods.

(c) We have obtained an optimal value for the interaction factor (λ) which minimizes the variance of the output response (for n=2)

$$\lambda_0 = 0.5$$

(d) For the special version of the general problem (model IV) where there are two subsystems with the same operational parameters we have concluded that large values (near the unity) of the interaction factor (λ) causes large fluctuations.

Hence if we aim to smooth results with short fluctuations we should restrict the magnitude of the interaction factor (near to zero).

Chapter 6

In chapter 6 we have examined carefully the mechanisms and philosophy underlying the PAYG funding method obtaining the following results:

(a) The existence of a contingency fund which absorbs fluctuations in mortality around a long term trend and fluctuations in fertility is the best vehicle in order to obtain inter-generational equity.

(b) The "smoothing" problem for the contribution rates and age of retirement is quite complicated in the general form so a special discrete algorithm has been designed which also requires a lot of simulation work.

(c) For the projected population of Greece the optimal path starts at (15%, 65) in year 1998 and ends at (16.5%, 67.5) in year 2020. Obviously the optimal path requires larger increases for the retirement age rather than for the contribution rate.

7.3 Perspectives & Future Research Topics

At this point, and after the description of the important results of this thesis, it is very interesting to suggest the potential perspective for future research. The proposals will be subdivided according to the chapter which refer to:

Chapter 4

(a) Examine potential modification in the model structure or insert different control actions (e.g. integral actions) in order to increase the critical value f_∞ (perhaps up to infinity) and so obtain a stable system even for large values of the delay factor (f).

(b) The result for the random input signal should be checked further using another full set of simulations (perhaps above 1000 simulations) in order to establish the existence of a minimum variance when (ε) takes value in the area of 50% - 60%.

(c) Other different random input signals may be considered except the uniform distribution (e.g. A random signal with an exponential distribution).

(d) For the time-varying model, it should be checked the transient behavior of the system under different patterns for (R) (e.g. R may be a random variable with a normal distribution).

Chapter 5

(a) Exercise simulations with typical values of the parameters involved in the problem (Λ matrix, $\underline{R}, \underline{e}, \underline{\varepsilon}$ and $\underline{F}$ vectors) and determine the transient and ultimate behavior of the system.

(b) Determine the condition for the area of equilibrium points (potential stability power) when some of the feedback factor $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ equals to zero.

(c) Considering different input signals (claims) design the optimal control for the premium strategy in order to minimize the variance of the surpluses and premiums.

Chapter 6

(a) In equation (6.10.2), we may include a stochastic element corresponding to random fluctuations in mortality & fertility patterns (the model will then become extremely complicated).

(b) Consider variable weights (i.e. $\theta(t)$ and $1-\theta(t)$) in equation (6.10.3). Normally a change in the age of retirement between ages 65 and 66 should be weighted less than a change between ages 68 and 69.

(c) Use variable values for Δc and Δr, according to the fund value for the discrete type of model [D-D]. For example when the accumulated fund is near to zero consider small changes (Δc and Δr) but when the fund is large (negative or positive) then consider bigger changes (Δc and Δr) for the contribution rates and age of retirement.

Closing our proposals we also suggest that future research may be directed to learning systems.

A learning system is a higher-level system than an adaptive control system. Quoting from Ogata (1970) we have the following description: "The approach to the design of such a system is to "teach" the system the best choice for each situation. Once the system has learned the optimal control law for each possible situation it may operate near the optimal condition regardless of the environmental changes".

Appendices

I. Z transformation (Laplace) [see Ogata (1970)]

A very useful and powerful tool for handling the complicated calculations of a system of difference equations is the Z-transofmation (Laplace transformation for differential equations).

Generally speaking, Z-transformation is a function between functions (i.e. matches a function with another unique function changing the free variable, $x_n \rightarrow x_z$).

Let us assume $\{x_n, n \in \mathbf{N}\}$ a sequence of real numbers. Then we produce a function of z-variable with the following formula

$$x_z = \sum_{n=0}^{\infty} x_n \cdot z^{-n} \quad \text{where} \quad |z| > 1$$

(the above series converge since $|z| > 1 \rightarrow \left|\frac{1}{z}\right| = |z^{-1}| < 1$) and we say that x_z is the Z-transformation of x_n. i.e. $\mathcal{Z}\{X_n\} = X_z$ and consequently $\mathcal{Z}^{-1}\{x_z\} = x_n$.

A simple example!

We consider the following sequence of numbers (known as a spike).

$$x_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \quad \text{i.e. } x_n = \{1,0,0,0,\ldots,0,\ldots\}$$

The Z-transformation is calculated below:

$$\mathcal{Z}\{x_n\} = \sum_{n=0}^{\infty} x_n \cdot z^{-n} = 1 \cdot z^0 + 0 + z^{-1} + 0 \cdot z^{-2} + \ldots = 1$$

We should give another equivalent version of Z-transformation which is calculated by the formula

$$x_z = \sum_{n=0}^{\infty} x_n \cdot z^n \quad \text{where} \quad |z| < 1.$$

It is easily proved that the slight modification of the formula does not affect the behavior or the properties of Z-transformation. We shall use either the first of the second definition for the solution of our problems.

We shall provide a table with the basic properties which are necessary for the development of our problems.

Table (I.1)

	x_n	x_z
1	$a \cdot x_n$	$a \cdot x_z$
2	$x_n + y_n$	$x_z + y_z$
3	x_{n+1}	$z \cdot x_z - z \cdot x_0$
4	x_{n+m}	$z^m x_z - z^{m-1} x_0 - \ldots - z x_{m-1}$
5	$x(\infty)$	$\lim_{z \to 1}\{(z-1) \cdot X_z\}$

The analogous concept in the continuous form is the Laplace transformation let f(t) a continuous functions of t then

$$L\{f(t)\} = F(s) = \int_0^{\infty} e^{-st} \cdot f(t) dt$$

II. Basics of Linear Algebra [see Kalogeropoulos (1995)]

This appendix provides a full reference for the calculation of the powers of matrix A.

Definition (II.1): Let $A \in C^{n\times n}$, $\lambda_i \in C$, $\underline{u}_i \in C^n$, then λ_i is called eigenvalue of A and u_i, ($\underline{u}_i \neq \underline{0}$) the respective eigenvector if and only if (by definition)

$$A\underline{u}_i = \lambda_i \underline{u}_i \leftrightarrow (A - \lambda_i I)\cdot \underline{u}_i = 0$$

Definition (II.2): The polynomial $\varphi(\lambda) = \det(\lambda I_n - A)$ is called the characteristic polynomial of A. $\varphi(\lambda)$ may be expressed as $\varphi(\lambda) = (\lambda - \lambda_1)^{\tau_1} \cdot (\lambda - \lambda_2)^{\tau_2} \cdot \ldots \cdot (\lambda - \lambda_\rho)^{\tau_\rho}$ with $\sum_{i=1}^{\rho} \tau_i = n$ and $\lambda_i \neq \lambda_j$ when $(i \neq j)$.

Definition (II.3): The set $\sigma(A) = \{\lambda_i \in C: \ (\lambda) = 0\}$ is called fasma of A and the integer number τ_i is called algebric manifold of λ_i.

Definition (II.4): The space $N_i = \{\underline{u}_i \in C^n : (A - \lambda_i I_n)\underline{u}_i = 0\}$ is called right «zero-space» of $A-\lambda_i I_n$ and d_i=n-rank($A-\lambda_i I$)=dim Ni is called geometric manifold of λ_i.

Definition (II.5): Let $A \in C^{n\times n}$ and $\lambda_i,\ldots,\lambda_\rho$ the eigenvalues of A. Then λ_i is called eigenvalue with simple structure if and only if $\tau_i=d_i$ (given $d_i \leq \tau_i$). If all the eigenvalues have simple structure then A is called matrix of simple structure.

Theorem (II.1): Let $A \in C^{n\times n}$, A may be represented by a diagonial matrix if and only if it is of simple structure i.e.

$$A=Q\cdot D\cdot Q^{-1}$$

where $D=\text{diag}\{\lambda_1,\lambda_2,\ldots,\lambda_\rho\}$

each eigenvalue appears at its algebric manifold indicator and $Q=[\underline{u}_1\underline{u}_2...\underline{u}_3]$ the matrix of eigenvectors.

Definition (II.6): Let $\underline{u}_k \in C^n$, $u_k \neq 0$ then $\underline{u}_k$ is called generalized eigenvector of k^{th} order if and only if

$$(A-\lambda I)^k \underline{u}_k = 0 \quad \text{and} \quad (A-\lambda I)^{k-1} \underline{u}_k \neq 0$$

Theorem (II.2): Let $A \in C^{n \times n}$, A may always be represent by the Jordan canonical matrix form

$$A=Q\cdot J\cdot Q^{-1}$$

where $J = \begin{bmatrix} J_{\lambda_1} & & & \\ & J_{\lambda_2} & \mathrm{O} & \\ & \mathrm{O} & \ddots & \\ & & & J_{\lambda_p} \end{bmatrix}$ and J_{λ_i} Jordan block

i.e. $J_{\lambda_i} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & \mathrm{O} & \\ & & \lambda_i & 1 & \\ & \mathrm{O} & & \ddots & \ddots \\ & & & \lambda_i & 1 \end{bmatrix}$ and $Q=[\underline{u}_1\underline{u}_2,...,\underline{u}_n]$

where $\underline{u}_i$, i=1,2,...,n the set of simple and generalized eigenvectors of the eigenvalues $\lambda_1,...,\lambda_p$.

Proposition (II.1): Let $A \in C^{n \times n}$ then $A_n=Q\cdot J^n\cdot Q^{-1}$ and more generally $f(A)=Q\cdot f(I)Q^{-1}$

Proof: $$A^n = (Q\cdot J\cdot Q^{-1})^n = \underbrace{(QJQ^{-1})(Q\cdot J\cdot Q^{-1})...(QJQ^{-1})}_{n-\text{times}} =$$

$$= QJ\underbrace{(Q^{-1}Q)}_{I_n}J\underbrace{(Q^{-1}Q)}_{I_n}...\underbrace{(Q^{-1}Q)}_{I_n}J\cdot Q^{-1} = Q\cdot \underbrace{J\cdot J...JQ^{-1}}_{n+m} = Q\cdot J^n\cdot Q^{-1}$$

III. Root Locus Method [see Shinners (1964)]

The root locus method has been developed (initially by Evans) in order to provide the ability of sketching the roots of the characteristic equation of a dynamic system with respect to a certain parameter involved. That parameter may be determined by the engineer of the system so the right choice will result a stable system. Ideally the engineer would like to know how the roots of characteristic polynomial move in the z-plane as the specific parameter changes from zero to infinity. The required graphical representation may be provided by the root locus method.

Actually the method relates the poles of the transfer function (or roots of the characteristic equation) of a closed loop system to the zeros and poles of the transfer function of the respective open-loop system. We have shown the existing relation (between the transfer function) in the section (2.5) i.e.:

$$G'(z)=\frac{G(z)}{1+G(z)H(z)} \tag{III.1}$$

Now let us describe the sketching procedure step by step quoting the rules from Shinners (1964) (p. 145-153) with small modifications and leaving the justification to the reference book.

[of course, the following procedure may be applicable to any equation

$$f(z)=0 \tag{III.2}$$

and not just for a characteristic equation].

Step 1: Rearrange equation (III.2) in the following form:

$$f(z)=1+K\cdot P(z)=0 \tag{III.3}$$

(So actually, if f(z) is the characteristic equation of a closed-loop system, P(z) is the transfer function of the respective open-loop system).

Step 2: Determine the number of separate loci which equals to the number of roots of the equation (III.2) or equivalently equals the order of the polynomial f(z) (characteristic polynomial).

Step 3: Determine the zeros and poles of equation

$$P(z) = 0 \tag{III.4}$$

keeping in mind that poles define the start of the root locus (k=0) while zeros define the end (k=∞) (if there are less finite poles than zeros then the root locus starts from infinity while in the opposite situation the root locus diverges to infinity.

(P(z) may be written in the form

$$P(z) = \frac{Q(z)}{R(z)} \tag{III.5}$$

so the poles of P(z) are the zeros of R(z)).

Step 4: Design the complex portion of the root locus keeping a symmetry along the x-axis as the complex roots or poles always occur as conjugate pairs (symmetrical to x-axis).

Step 5: Design the real portion of root locus keeping in mind that these portions always lie to the left of an odd number of finite poles plus finite zeros.

Step 6: Design the asymptotes of the root loci by finding firstly the angles $(ang)_n$ with x-axis i.e.

$$(\text{ang})_n = \pm\frac{n\pi}{p-z} \qquad \text{(III.6)}$$

where $\quad p$ = number of finite poles of $P(z)=0$

z = number of finite zeros of $P(z)=0$

n = an odd integer (1,3,5,7,...)

Step 7: Finalize the designation of the assymptotes by finding the intersection point with x-axis.

$$(\text{IS}) = \frac{\sum(\text{values of poles}) - \sum(\text{values of zeros})}{(\text{Number of finite poles}) - (\text{Number of finite zeros})} \qquad \text{(III.7)}$$

Step 8: Determine the points of breakaway of the root locus from the real axis. The determination of these points requires the solution of the equation below.

$$\frac{d}{dz}\left[\frac{-1}{P(z)}\right] = 0 \qquad \text{(III.8)}$$

Of course, we are going to accept only the solutions lying in the expected areas.

Step 9: Determine the angles made by the root locus leaving a complex pole (if any). This can be done using the angle criteria described at the end of this appendix.

Step 10: Determine the intersection of the root locus with some critical lines i.e. the imaginary axis or the unit circle etc.

After the completion of the ten steps we may sketch with enough accuracy the loot locus in the z-plane.

Finally and before closing the appendix we shall state two very useful criteria in the exploration of the root locus.

As the P(z) polynomial is a complex function may take complex values so for a specific point z* which lies on the root locus the following relation holds:

Magnitude Criteria: $K = \frac{1}{|P(z)|}$ (III.9)

Proof: Considering equation (III.3) we obtain that

$$1 + k \cdot P(z) = 0 \Rightarrow K \cdot P(z) = -1 \Leftrightarrow$$

$$|k \cdot P(z)| = |-1| \Leftrightarrow K \cdot |P(z)| = 1 \Leftrightarrow$$

$$K = \frac{1}{|P(z)|}$$

Angle Criteria: $\measuredangle\, P(z) = (2i + 1)180^\circ \quad i=0,1,2,\ldots$

Proof: From (III.3) again and taking the angles we obtain

$$\measuredangle\, k \cdot P(z) = \measuredangle -1 \Leftrightarrow$$

$$\measuredangle k + \measuredangle P(z) = \measuredangle -1 \Leftrightarrow$$

$$0 + \measuredangle P(z) = (2i+1)180^\circ \quad i=0,1,2,\ldots$$

IV. Optimization Criteria in Standard and Functional Analysis problems [see Athans & Falb (1966)]

(1) Firstly, consider a real valued function which is defined also in **R**: i.e. $f:(a,b)-\{x_1,\ldots,x_n\}\rightarrow R$ then a minimum may exist.

(a) At the boundaries of the (a,b) i.e. a or b or at the points of [a,b] where the f is not defined $x_1,x_2,\ldots,x_n$. So we have to examine $f(a), f(b), f(x_1), \ldots, f(x_n)$.

(b) At the points where the 1st derivative does not exist.

(c) At the points where the 1st derivative exists and equals zero

$$f'(x^*)=0$$

then x* is a minimum if and only if the second derivative at this point is positive (i.e. $f''(x^*)>0$).

(2) We shall continue with a real valued function which is defined in $\mathbb{R}^n$ i.e.

$$f:\Omega\subseteq\mathbb{R}^n\rightarrow\mathbb{R} \qquad \text{then a minimum exists}$$

(a) ... the same as (1)

(b) ... the same as (2) substituting the concept of derivative with the gradient of the function

(c) At the points where the gradient exists and equals zero (all the partial derivatives equals to zero) i.e.

$$\nabla f=\underline{O}\Leftrightarrow\frac{\partial f}{\partial x_1}(x^*)=\frac{\partial f}{\partial x_2}(x^*)=\ldots=\frac{\partial f}{\partial x_n}(x^*)=0$$

Then x* is a minimum if the matrix

$$Q = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_1 x_n}(x^*) \\ \vdots & \vdots & & \\ \frac{\partial^2 f}{\partial x_n \partial x1}(x^*) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x^*) \end{bmatrix}$$

is positive definite.

A matrix $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$ is said to be positive definite if and only if

the minors

$$p_1 = a_{11}, p_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \ldots, p_n = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

are all greater than zero (for a semidefinite, may equals zero).

(3) Ordinary minima with Constraits (Lagrange multipliers).

(We shall quote the respective theorem from Athans & Falb (1966) p. 235).

If x* is a point of the domain D at which f has a local minimum subject to the constraits $g_i=0$, $i=1,2,\ldots,m$ where f and the g_i all have continuous partial derivatives then there are numbers $p_1^*, p_2^*, \ldots, p_m^*$ called Lagrange multipliers such that $x^*, p_1^*, p_2^*, \ldots, p_m^*$ are a solution of the following system of n+m equations in the n+m unknows

$x_1, x_2, \ldots, x_n, p_1, p_2, \ldots, p_m$:

$$\frac{\partial f}{\partial x_i} + p_1 \frac{\partial g_1}{\partial x_i} + p_2 \frac{\partial g_2}{\partial x_i} + \ldots + p_m \frac{\partial g_m}{\partial x_i} = 0$$

$$g_i(x_1, x_2, \ldots, x_n) = 0$$

where i=1,2,...,n and j=1,2,...,m.

(4) Finally we shall briefly discuss the functional minima (more analysis is provided in Athans & Falb (1966) p. 237-363 from where we quote the result below).

The problem has a variety of forms. Generally speaking, it is a problem of finding the optimal function control which minimizes a certain criterion (usually given as a functional) under the contraight of a differential equation.

Let us consider the dynamic equation $\dot{x}(t) = f(x(t), u(t))$ then we require the optimal u*(t) which minimizes L i.e.

$$\min_u L = \min_u \int_{t_0}^{t_1} L(x(t), u(t))dt$$

given that $x(t_0) = x_0$ and $x(t_1) = x_1$.

We must introduce the Hamiltonian H of the system

$H(x(t), p(t), u(t), t) = L(x(t), u(t)) + \langle p(t), f(x(t), u(t)) \rangle$

or $H(t) = L(x(t), u(t)) + p(t) \cdot f(x(t), u(t))$ then $H(x^*(t), p^*(t) u, t)$ has an absolute minimum as a function of u when $u = u^*(t)$ $t \in [t_0, t_1]$ i.e.

$$\frac{\partial H}{\partial u} = 0 \quad \text{or} \quad \frac{\partial H(t)}{\partial u(t)} = 0 \quad \forall t \in [t_0, t_1]$$

while $$p^*(t) = -\frac{\partial H(t)}{\partial x(t)}, \quad x'(t) = \frac{\partial H(t)}{\partial p}$$

References

- Ackman R.C., Green P.A.G., Young A.G. (1985). "Estimating Claims Outstanding". Actuarial Education Senate, Oxford.
- Athans M. & Falb P.L. (1966). "Optimal Control" McGraw-Hill Book Company, New York.
- Balzer L.A. & Benjamin S. (1980). "Dynamic response of insurance systems with delayed profit/loss sharing feedback to isolated unpredicted claims". Journal of the Institute of Actuaries, vol. 107, 513-528.
- Balzer L.A. (1982). "Control of insurance systems with delayed profit/loss sharing feedback and persisting unpredicted claims". Journal of the Institute of Actuaries, voll 109, 285-316.
- Benjamin S. (1984). "An actuarial layman looks at control theory". Transactions of 22nd International Congress of Actuaries. Sydney 1984 vol. 4, 295-310.
- Benjamin S. (1989). "Driving the Pension fund". Journal of the Institute of Actuaries vol 116, 717-735.
- Borch K. (1967). "The theory of risk. Journal of the Royal Statistical Society Ser B29, 432-452.
- Bowers N.L., Gerber, H.U., Hickman, J.C. & Nesbitt, C.J. (1986). "Actuarial Mathematics" The Society of Actuaries Illinois.
- Brown A. & Grenfell C. (1992). "Crediting rate management for Capital guaranteed insurance". Transactions of the 24th International Congress of Actuaries, 15-34.
- Brown R.L. (1992). "Pay-as-you-go funding stability. An age of eligibility model". Transactions of the 24th International Congress of Actuaries p. 35-56.

- Buhlmann H. (1997), "The actuary: the role and limitations of the profession since the mid-19th century", ASTIN Bulletin vol. 27, No 2, pp 165-171.
- Butz W.P. & Ward M. (1979), "The emergence of counter-cyclical U.S. Fertility", American Economic Review 69 (June), p. 318-28.
- Chen Yung-Ping (1987): Making Assets out of Tomorrow's Elderly. The Gerentologist 27(4), p. 410-16.
- Daykin C.D. (1998). "Can we avert the old-age crisis". Trans. of 26th International Congress of Actuaries, 415-442.
- Daykin C.D. & Pentikäinen & Pessonen M. (1994). "Practical Risk theory for Actuaries" Chapman & Hall (sections 6.6.3., 6.6.4.).
- De Finetti B. (1957). "Su una impostazione alternativa della theoria collectiva del rischio". Transactions of the 15th International Congress of Actuaries. New York Vol. II, p. 433-443.
- Dilnot A., Disney R., Johnson P. & Whitehouse E. (1994): "Pension Policy in the U.K.: An economic Analysis". The Institute for Fiscal Studies, London.
- Earlstein R.A. (1987), "Birth and Fortune, the Impact of Numbers on Personal Welfare, London Grand McIntyre.
- Ermisch J.F. (1983), "The political economy of Demographic Change, Policy Studies Institude, London, Heinemann Books.
- Fujiki Masa-Hiko (1994). "Pension Fund Valuation" Ph.D. Thesis, City University.
- Gadzow, J.A. (1973). "Discrete-time systems" Prentice-Hall, Inc, Englewood Cliffs, New Jersey.
- Gerber H.U. (1979). An introduction to Mathematical Risk Theory S.S. Huebner Foundation, University of Pensylvania Chapters 5,7.

- Goldberg S. (1963). "Introduction to Difference Equations".
- Greek National Statistical Service (1995), Tables 1995.
- Haberman S. & Sung J.H. (1994), "Dynamic Approaches to Pension Funding". Insurance Mathematics and Economics vol. 15, p. 151-162.
- Kalogeropoulos G. (1995). "Matrix theory and applications". Lecturer Notes University of Athens.
- Kalogeropoulos, G. (1987). "Lecture notes on Control theory" University of Athens.
- Kamano S. & Nara S. (1992). "Long-term management of reserve and dividend". Transactions of the 24th International Congress of Actuaries, 123-129.
- Kamien M.I & Schwartz N.L (1981), "Dynamic optimization: the calculus of variation and optimal control in Economics and Management", North Holland, New York.
- Keyfitz N. (1985), "Applied Mathematical Demography" (2nd edition) Springer Verlag, New York.
- Keyfitz N. (1985), "The demographics of unfunded pensions", European Journal of Population 1, p. 5-30.
- Lapkoff S.F. (1991). "A research note of keyfitz's. The demographics of unfunded pensions". European Journal of Population 7, 159-169.
- Loades D.H. (1992). "Instability in pension funding". Transactions of 24th International Congress of Actuaries, Vol. 2, 137-154.
- Loades D.H. (1998). "Elementary Engineering Control and Pension Funding". Trans. of the 26th International Congress of Actuaries, 239-266.
- Martin - Lőf A. (1983). "Premium control in an insurance system, an approach using linear control theory. Scandinavian Actuarial Journal, 1-27.

- Martin - Lőf A. (1994). "Lectures on the use of Control theory in insurance". Scandinavian Actuarial Journal vol. 1, 1-25.
- Mc Cutcheon & Scott (1986), "An Introduction to the Mathematics of Finance", Institute of Actuaries & the Faculty of Actuaries, Butterworth Heinemann.
- Nesbitt C. (1991). "Elementary models of reserve fund of the old age, Survivors & disability insurance (OASDI) in the USA". Actuarial Research Clearing House (ARCH) p. 63-91.
- Nesbitt C. Rosenberg M., Berger L., Levinsky M., Nerdum A., O' Donnell S., Samaniego D., Trendel A. (1995). "Some financing options for Social Security". Actuarial Research Cleaning House (ARCH), 1, p. 381-417.
- Nesbitt C., Nerdum A. & Clark S. (1995). "Conclusions from Michigan Studies of Social Security Financing".
- Neuhaus W. (1989). "Mutual reinsurance and homogeneous linear estimation" ASTIN Bulletin vol 19, 213-221.
- O' Brien T. (1987). "A two-parameter family of pension contribution function and stochastic optimization". Insurance Mathematics & Economics pp 129-134.
- Ogata K. (1970). "Modern Control Engineering" Prentice Hall, Inc, Englewood Cliffs N.J.
- Pesonen M.J. (1984). "Optimal Reinsurance". Scandinavian Actuarial Journal pp 65-90.
- Rantala J. (1986). "Experience Rating of ARIMA Processes by the kalman Filter". Astin Bulletin 16, 19-31.
- Rantala J. (1987). "On experience rating and optimal reinsurance". ASTIN Bulletin vol 19, No 2, 153-178.

- Rantala J. (1988). "Fluctuations in Insurance business results. Some control theoretical aspects". Transactions of 23^{rd} International Congress of Actuaries, 43-121.
- Samuelson, P. (1958), "An exact consumption - loan model of interest with or without the social contrivance of money", Journal of Political Economy 66(6), 467-482.
- Seal H.L. (1970). "Simulation of the ruin potential of non-life insurance companies". Transactions of the Society of Actuaries, Vol. 21, 563-590.
- Shinners S.M. (1964). "Control system design. John Wiley & Sons, Inc, New York. (Root locus method, p. 743-153).
- Sung J. (1997). "Dynamic Programming Approaches to pension funding" Ph-D thesis. The City University, London.
- Taylor G.C. (1987). "Control of unfunded and partially funded systems of payments". Journal of the Institute of Actuaries, vol. 114, 371-392.
- Trowbridge C.L. (1963). "The unfunded Present Value Family of Pension funding Methods". Transactions of the Society of Actuaries, 15 pp. 151-169.
- Trowbridge C.L. (1989). "Fundamental Concepts of Actuarial Science "The Society of Actuaries, Chapter VIII (Assumptions Conservatism and Ajustment).
- Trowbridge J.R. (1977). "Assessmentism - An Alternative to pension funding". Journal of the Institute of Actuaries vol 104, 173-204.
- U.S. Department of Commerse (1987), p. 46-62.
- Vandebroek M. & Dhaene J. (1990). "Optimal premium control in a Non-Life Insurance Business" Scand. Actuarial Journal, 3-13.
- Vandebroek M. (1990). "Pension funding and Optimal Control". Mitteilungen der Vereinigung schweizericher Versicherungsmathematiker 313-325.

- Waters H.R. (1986). "Introduction to Credibility Theory". Actuarial Education Senate, Oxford.
- Wilkie A.D. (1997). "Human genetics - uncertainties and the financial implications ahead". The Royal Society and the actuarial profession Discussion meeting at the Royal Society.
- William M. Mercer International (1988). "Multinational pooling. The Networks".
- Willekes F. & van den Hoogen H. (1998). "Solidarity Assumed: Solidarity in Life insurance". Trans. of International Congress of Actuaries, pp. 147-178.
- World Bank (1994). "Averting the old-age crisis". Oxford University Press.
- Zimbidis A. & Haberman S. (1993). "Delay, feedback and variability of pension contributions and fund levels". Insurance Mathematics and Economics vol 13, 271-285.