PROVISIONAL MILITARY GOVERNMENT
OF SOCIALIST ETHIOPIA
MINISTRY OF AGRICULTURE

APPLICATION OF TIME SERIES ANALYSIS IN AGRICULTURE

STATISTICS SECTION
PLANNING AND PROGRAMMING DEPARTMENT

Addis Ababa, February 1980.

CONTENTS

ZR

PREFACE

In recent years the wide spread applicability of time series analysis as a tool for modelling dynamic systems, and forecasting the behaviour of the output of such systems has given it a better recognition and importance.

In particular, time series analysis has a practical application in explaining the different structures of agriculture: crop production, area under different crops, yield of different crops, prices of agricultural products, marketed production, cost of crop production, livestock, consumption of agricultural products, income and expenditure of rural population, nutrition, etc. These type of data are the most important source of information on the characteristics features of the development of agriculture and on the changing relations among economic, social and natural phenomena.

In Ethiopia, the period for which we have time series data on agriculture is very short. We have time series data on: crop production, area under different crops and yield of different crops for the last 6 years (1974/75-1979/80); prices of agricultural products (retail price) for the last 20 years and cost of crop production for three years. However, the new FAO/UNDP project:"Integrated System of Food and Agricultural Statistics", which will be launched in July 1980, will conduct agricultural surveys regurally to develop a system of statistical analysis for planners and other users. Therefore, it is necessary to start developing and applying time series analysis in our agricultural problems for it is one of the important statistical methods used for these purposes.

There are two approaches to statistical analysis of time series data. First, it is analysis of the behaviour of singular time series in order to assess its present and possible future development. The second approach is analysis

of interpretations between many time series variables explored through the simple and multiple regression analysis.

In principle, the main topic of this publication is univariate time series methods. By univariate analysis we mean description of the historical and future time series data in terms of its previous values and/or in terms of mathematical funtions of time variable t. Such methods seem to be highly useful in agricultural analysis. The univariate approach provides the methodology for analysing of trends and measures of periodical variation in time series with short-term (monthly or quarterly) observations and for forecasting of time series with monthly, quaterly and annual observations. Univariate time series methods may be applied also to data showing changes of agricultural production, yield and area under crops by region, i.e. to time series cross-section data. The advantage of this approach is that to perform any forecast, it doeas not require supplementary data on other phenomena infuencing the forecasted one.

This publication consists of three chapters. Chapter I contains the basic information on the linear correlation and regression between two variables and on the estimation problem necessary for explaining the topics covered in the remaining chapters.

In chapter II, the methods of time series decomposition are presented. The method of seasonal adjustment may prove to be useful for analysis of data on market and wholesale prices of agricultural goods. Analysis of actual trends of prices as well as their seasonal variation may serve as an auxiliary tool for the assessment of food situation at wereda, awraja and regional level.

The different methods of forecasting are described in chapter III, and they may prove particularly useful in forecasting crop production, yield of different crops and area under these crops at the country and regional levels. Of course, in some fields of statistical analysis, regression ana-

lysis methods, particularly multiple regression ones, are irreplaceble. Econometric models are particularly useful when we would like to reveal or know more precisely the relationship among economic, social and natural phenomena, and when we want to estimate the possible answer of investigated variable on the change of other variables. The introductory information on the multiple regression models and hypothetical models for crop production and area under crops are included in chapter III as its final part.

This publication was prepared by Dr. Stefan Giembicki, FAO Consultant in Time Series Analysis, under FAO/UNDP Project ETH/73/004. The author spent 9 months in Ethiopia in 1979, during which he gave series of lectures on time series analysis methods for the participants from different institutions. He has been working in the Research Centre of the Central Statistical Office, Poland, since 1966 ,and has published a number of papers on time series analysis.

We hope that this publication will be useful for the users of agricultural statistics. Any comments that would help us to improve the quality of our surveys and analysis are most welcome.

Ghebre-Selassie Mebrahtu

Addis Ababa, December 1980

Head of the Statistics Section

Planning and Programming Department

CHAPTER I

LINEAR ASSOCIATION BETWEEN TWO VARIABLES

1.1. INTRODUCTION

Economical and statistical analysts are very often interested in the relationship between pairs of variables. To understand such relationships, we should distinguish between two different kinds of question which could be asked.

1) Is there any evidence of relationship (or association) between X and Y?

2) If we are able to specify the form of the relationship between X and Y (for example a linear relationship $Y = a + bX$) how we can estimate the parameters in the relationship from a sample of observations?

The answer to the first question is explored through correlation analysis. The second type of question is dealt with regression analysis. We confine ourselves to the simplest case of the association between two variables X and Y.

It should be noted, that the first question, above, does not specify any form for the relationship between X and Y. In particular it does not specify the direction of causality between Y and X. This is important point, that the correlation analysis techniques are designed to measure the degree of association between X and Y and they cannot be used to prove that a causal link exists between X and Y.

In discussing the association, we shall consider the case in which X and Y are associated in a linear fashion. There are more general tests for association which does not assume linearity.

1.2. THE CORRELATION COEFFICIENT

To illustrate development of a measure of association we shall analyse the following (hypothetical) data. We have data on the use of fertilizers (in kilograms per hectare) and yields of wheat (in quintals per hectare) from the random sample of 20 farmers.

Table 1

Use of fertilizers and yield of wheat for a sample of farms

i	Use of fertilizers (X_i)	Yield of wheat (Y_i)	i	Use of fertilizers (X_i)	Yield of wheat (Y_i)
1	10	8.7	11	2	7.5
2	15	16.0	12	0	8.4
3	5	10.1	13	1	6.2
4	3	7.6	14	8	19.0
5	11	17.0	15	7	13.0
6	10	18.5	16	4	8.9
7	14	26.0	17	12	15.0
8	20	29.0	18	9	18.0
9	16	27.0	19	21	31.0
10	6	15.5	20	13	22.0

As the first step in the analysis we may examine the data visually by plotting them in a scatter diagram. This is done in Fig. 1, where there does appear to be some positive association between the level of applied

fertilizers and yields of wheat, because for bigger amount of fertilizers corresponds in general, higher yield of wheat.

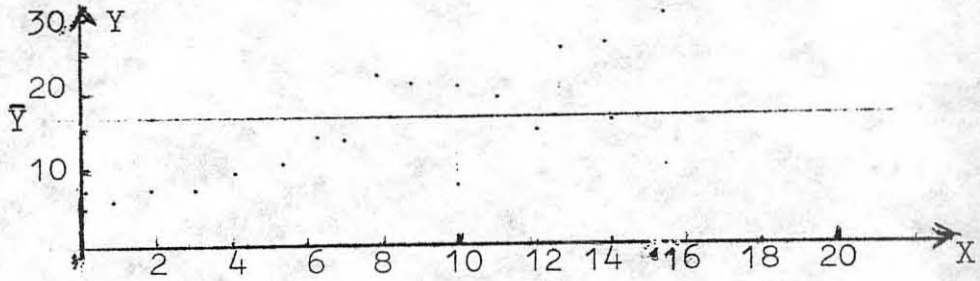


Fig. 1.
Scatter diagram of data from Table 1.

We may state that points in Fig. 1 are dispersed and that positive association between Y and X is association between two dispersed variables. It means that each value of Y variable corresponds not with one and only one value of Y variable but also with some sub-sample of Y variable observations. Similarly each value of Y variable may be accompanied by a set of X variable values.

Dispersion of the statistical variable is measured as dispersion around its mean value. In Fig. 1 a straight line parallel to X axis intersect Y-axis at the point

$$\bar{Y} = \frac{1}{20}\sum_{i=1}^{20} Y_i = 16.2 \text{ qt/ha} \qquad (1.1)$$

which is arithmetic mean of variable Y.

Sample variance

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{20}(Y_i-\bar{Y})^2 \qquad (1.2)$$

shows what is average squared distance between observations of the variable Y and its mean value calculated on the base of a sample.

Sample standard deviation

$$S_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{20}(Y_i - \bar{Y})^2} \qquad (1.3.)$$

shows average dispersion of the Y variable observations about its mean calculated for the sample of observations.

To obtain a numerical measure of the degree of linear association we proceed as follows:

1) Calculate the arithmetic means of the variables, (obtaining $\bar{X}$ = 13.25 kg/ha and $\bar{Y}$ = 16.2. qt/ha).

2) Transform our observations to deviations from the sample means defining $x_i = X_i - \bar{X}$, and $y_i = Y_i - \bar{Y}$, $i = 1, 2, \ldots, 20$. The new variables x_i and y_i may be illustrated geometrically as in Fig.2.

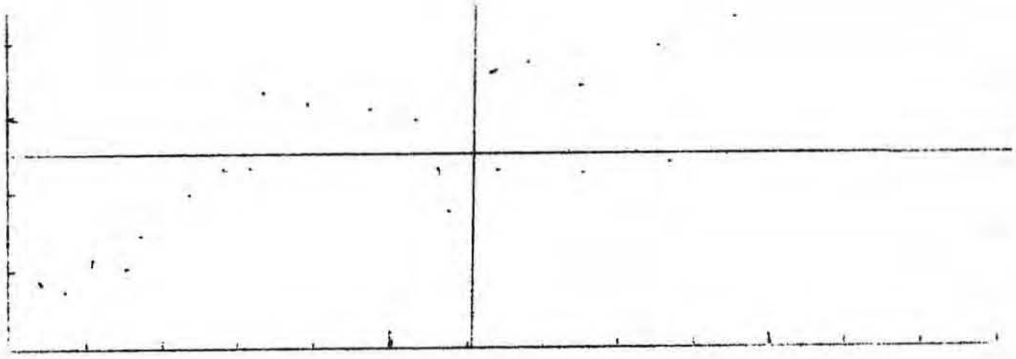


Fig.2. Scatter diagram of deviations from the sample means (x_i , y_i).

In Fig.2 the origin of the axis is point $(\bar{X}, \bar{Y})$. The (x_i , y_i) points are distributed between the four quadrants of the $\bar{X}$, $\bar{Y}$ axes. These quadrants we shall number I through IV.

Since x_i and y_i are deviations from the (sample) means they will take both positive and negative values depending on the quadrant in which they occur, as will the products $x_i\, y_i$. In quadrant I both x_i and y_i will be negative, so that the products $x_i\, y_i$ will be positive. In quadrant III, x_i and y_i are positive, so that the products $x_i\, y_i$ are positive.

On the other hand in quadrants II and IV either x_i or y_i will be positive (negative) when the other is negative (positive), so that the products $x_i\, y_i$ will be negative in this two quadrants.

The sign pattern of $x_i \cdot y_i$ suggests a numerical measure of association. If there is a positive association between x and y we would expect a high proportion of $(x_i\, y_i)$ points to fall in quadrants I and III, and hence a sum of products $x_i\, y_i$ will be positive. If there is a negative association between x and y we would expect most of the $(x_i\, y_i)$ to fall in quadrants II and IV and hence $\sum x_i\, y_i$ will be negative. If there is no association between x and y we would expect to find the $(x_i\, y_i)$ points scattered over all four quadrants, with positive and negative $x_i\, y_i$ products tending to cancel out so that $\sum x_i\, y_i$ will tend to be close to zero.

However $\sum x_i\, y_i$ is a rather crude measure of association because it depends not only on the strength of association but on the sample size too. We may counter

this by considering not $X_i\ y_i$ but its mean value

$$\frac{1}{N} \sum x_i\ y_i \qquad (1.4)$$

but this still leaves us with a verify problem of interpretation because the magnitude of $\sum x_i\ y_i$ depends on the units of measurement of the variables. For example $\sum x_i\ y_i = 800$ and $\sum x_i\ y_i = 8$ could actually refer to the same variable if in the first case x is measured in metric quintals when in the second in kilograms. The problem of units of measure may be solved if the average sum of products (1.4) is divided by the product of the standard deviations of the variables, S_x and S_y. The ratio which results from these operations is called the coefficient of correlation, r

$$r = \frac{\sum x_i\ y_i}{n\ S_X S_Y} \qquad (1.5)$$

where

$$S_X = \sqrt{\sum (X_i - \bar{X})^2/N} \qquad S_Y = \sqrt{\sum (Y_i - \bar{Y})^2/N}$$

so that

$$r = \frac{\sum x_i\ y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

For computation the following alternative way of expressing r is convienient

$$r = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N}}{\sqrt{\sum X_i^2 - \frac{(\sum X_i)^2}{N}} \sqrt{\sum Y_i^2 - \frac{(\sum Y_i)^2}{N}}} \qquad (1.6)$$

The coefficient of correlation is a pure mumber(without units) lying in the range $-1 \leq r \leq 1$, taking the values ± 1 when all the points lie exactly on a straight line, the sign depending on whether the line has a positive or negative slope.

For the negative association between x and y coefficient of correlatiom is negative ($r < o$) for the positive association, $r > o$.

For assessment, if correlation coefficient within (-1, 1) border shows significant association special tests were developed. However, it is not the subject of our work to present these tests. Readers interested in the topic may find it in manuals of statistics, for instance [1 , 4 , 10] .
We will confine ourselves to the statement that reliability of r measure depends, among others, on the size of sample. For small sample the calculated correlation coefficient value may be influenced by chance, differing strongly from the true correlation coefficient between variables X and Y.

The sample correlation coefficient is conceptually simple and is easy to compute, but its interpretation does require some care. In the first place, as has already been mentioned, the correlation coefficient measures the degree of linear association between X and Y. Thus, if we take r as non-significant we have not ruled out the possibility of a strong non linear relationship between x and y. The latter case must be investigated by other methods, the simplest of which is to plot the data in a scatter diagram. For example, the reader may plot x and y where $y = x^2$ and $x = -4$,

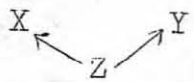
-3, -2, -1, 1, 2, 3, 4 and calculate the correlation between x and y. Tests for non linear association are presented among others in [4, 10].

The second problem which has also been mentioned already is the danger involved in interpretting a strong association as a causal relationship. A high correlation between X and Y does not prove that X causes Y or viceversa. Here we may distinguish two situations

I

$X \longleftrightarrow Y$

II



In situation I, X and Y are causally linked and we observe Y increasing (decreasing) because X is increasing (decreasing). In situation II, X and Y are not causally linked to each other, but both are positively linked to a third variable Z. Thus if Z increases, both X and Y increase and we observe a high correlation between them. This kind of correlation is named " spurious" correlation. Often, when the observation on X and Y are collected over time a good example of Z variable is the growth of population.

The one important thing here is that sometime not too much weight should be placed on a significant sample correlation coefficient in isolation, since to establish causal linkage one may require a considerable amount of further research.

1.3. SIMPLE REGRESSION ANALYSIS AND LEAST SQUARES METHOD

We turn now to the estimation of the parameters in a relationship between a two variables, i.e. some variable Y is a linear function of a single explanatory variable X. To illustrate the estimation of a linear re-

lationship consider the data on yield of wheat (Y) and use of fertilizers (X), that we reported in Table 1. We shall suppose that there may be a linear relationship.

$$Y_i = a + bX_i + e_i \qquad (3)$$

where errors e_i reflect the fact that we expect discrepancies between adjusted function and observations Y, because of another variables than X influence Y but are not included into model.

We have data on the X and Y but errors e_i are unobservable. Parameters a and b are unknown and our object is to estimate them from our data. The scatter diagram for these two variables is reproduced in Fig. 3 in which straight line labelled a + bX represents the (unknown) true relationship.

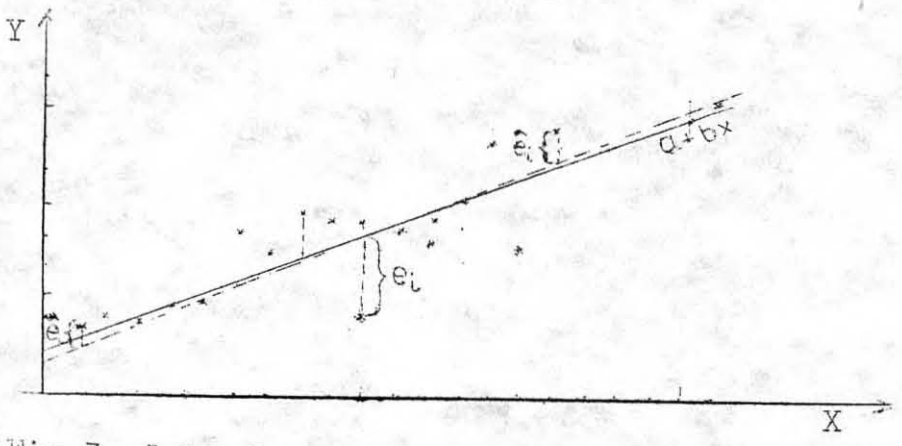


Fig.3. Data from Table 1, and a fitted straight line.

Points of this straight line determine the relation between X and Y variables. Our object is to draw some straight line which is an estimate of a(a+bX) line. Denote this fitted line as $\hat{Y} = \hat{a} + \hat{b}X$ (on fig 3 dotted line). Differences between the observations Y_s and fitted line

$$Y_i - \hat{Y}_i = \hat{e}_i \qquad (1.4)$$

are called residuals.

The simple effective and accurate method for the estimation of a and b parameters is method of least squares.

The method of least squares accomplishes two objectives:

1. The sum of the vertical deviations of the observed values from the fitted straight line equals zero. If a vertical lines were to be drawn, as in Fig.3, from each Y value to the straight line, the vertical lines extending upwards from the fitted line would exactly balance those extending downwards. This fitted line is not the only straight line from which the algebraic sum of the deviations equals zero; as a matter of fact, any straight line (other than vertical) which passes through $(\bar{Y}, \bar{X})$ point fulfils this requirement.
2. The sum of the squares of all these deviations is less than the sum of the squared vertical deviations from any other straight line or in other words the sum of the squares of this deviations is minimal. This is the reason why the method of fitting is called the method of least squares.

This requirement may be presented as follows:

$$D = \sum \hat{e}_t = \sum (Y_i - \hat{Y}_i)^2 = \text{minimum}$$

or

$$D = \sum (Y_i - a - bX)^2 = \text{minimum} \qquad (1.5)$$

The sum of squares D depends only on the adjusted values of a and b parameters (Y and X are data). Therefore we should find such a and b which will minimize the sum of squares of D.

From the mathematical analysis results that the necessary conditions for the function D to be minimized is

$$\frac{\partial D}{\partial a} = 0 \quad \text{and} \quad \frac{\partial D}{\partial b} = 0 \qquad (1.6)$$

where $\frac{\partial D}{\partial a}$ and $\frac{\partial D}{\partial b}$ are partial derivatives of function D with respect to a and b. These give the 2 normal equations for a and b, namely,

$$\sum Y_i = N + b\sum X_i$$
$$\sum Y_i X_i = a\sum X_i + b\sum X_i^2 \qquad (1.7)$$

solving this set of equations with respect to a and b we find:

$$b = \frac{\sum Y_i X_i - \frac{\sum Y_i \sum X_i}{N}}{\sum X_i^2 - \frac{(\sum X_i)^2}{N}}$$
$$a = \bar{Y} - b\bar{X} \qquad (1.9)$$

where $\bar{Y} = \frac{1}{N}\sum Y_i$, $\bar{X} = \frac{1}{N}\sum X_i$

$i = 1, 2, \ldots, N$

additionally we see that if we divide through the first equation in (1.7) by N we obtain

$$\bar{Y} = a + b\bar{X}$$

that is, the least squares estimates are such that the estimated line passes through the point of means $(\bar{Y}, \bar{X})$.

Using formulas (1.8), (1.9) we get the following relationship between yield of wheat and use of fertilizers in our example:

- parameters:

$\hat{a} = 5.5$ $\hat{b} = 1.14$

- regression equation

$$\hat{Y}_i = 5.5 + 1.14X_i \qquad (1.10)$$

- coefficient of correlation between X and Y is r=0.898.

Taking fitted function (1.10) as approximation of real relation between Y and X we may find for every value of used fertilizers i.e. for every value of X variable within its range of variability, corresponding yield of wheat.

1.5. THE STATISTICAL PROPERTIES OF THE LEAST SQUARES ESTIMATORS

Economic and social phenomena are stochastic by their nature. It means that every observation of economic or social variable consists of two parts: 1) real parts, i.e. level common for all observations and regular changes of the level, 2) individual features of observations or in other words random variability around the real part.

Regression model should include variables explaining generation of the real part of the variable excluding factors causing its individual (random) variability. This individual variability is included into regression equation as unobservable random error term.

It is assumed in regression analysis that error term should fulfil following conditions:

$$E(e_i) = 0 \qquad i = 1,2,\ldots,N \tag{1.10}$$

$$E(e_i^2) = \sigma^2 \tag{1.11}$$

$$E(e_i e_j) = 0 \qquad \text{when } i \neq j \tag{1.12}$$

where E is symbol of expected value (see [1], [4]). The first assumption is that e_i are random variables having zero means, implying that Y_s are randomly distributed around the true relationship so that $E(Y_i) = E(a + bX_i + e_i) = a + bX_i$. Assumption (1.11) implies that e_i represent random variables with the same variance σ^2 and assumption (1.12) shows that covariances between random variables e_i, e_j, $i \neq j$ are equal zero, then this variables are independent.

Additional assumption is that explaining variable X represents a set of constant values what implies that only source of random variability in variable Y are random errors e_i.

The parameters a and b estimated by means of least squares are linear, unbiased and have property of minimum variance (compare [10] , pages 137-140). The first property means that a and b are linear function of dependent variable Y. This propertly has practical implication namely it simplifies derivation of statistical measures of accuracy(compare [10] , pp.140-144). Second property means that estimated parameter may differ from true parameter only randomly. The last property is that out of the class of linear unbiased estimators of a and b, $\hat{a}$ and $\hat{b}$ have the smallest sampling variance. Least squares estimators have above properties even if some of the assumptions (1.11)-(1.12) are not fulfiled. If we impose additional condition for e_i terms, that random errors come from normally distributed random variables it is possible to develop tests for hypothesis about statistical significance of the parameters and to construct, limits within whose true parameters should be found with a high probability (compare [1] , [10]).

1.6. GOODNESS OF FIT

If assumptions about error term in regression equation are fulfiled it can be proved (see [10] , p.142) that residuals

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{a} - \hat{b}X_i \qquad (1.13)$$

are unbiased estimators of the errors e_i and their sample mean equals zero. Therefore it is shown (compare [10] , p.150) that unbiased estimator of error term variance σ^2 is

$$S_e^2 = \sum e_i^2/(N-2) \qquad (1.14)$$

Square root of S_e^2

$$S_e = \sqrt{\sum e_i^2/(N-2)} \qquad (1.15)$$

is called <u>standard error of estimate</u> and is applied as a measure showing average dispersion of the variable around fitted function. Standard error may be interpreted as average error which can be commited when we take approximated relation as a true one.

After substituting $a = \bar{Y}-b\bar{X}$ into (1.13), we get

$$\hat{e}_i = (Y_i - \bar{Y}) - b(X-\bar{X}) \qquad (1.16)$$

or $$\hat{e}_i = y_i - bx_i \qquad (1.17)$$

where $y_i = Y_i - \bar{Y}$, $x_i = X_i - \bar{X}$

taking square of $\hat{e}_i$ we receive

$$\hat{e}_i^2 = y_i^2 - 2by_ix_i + b^2x_i^2 \qquad (1.17)$$

since $by_ix_i = b^2x_i^2$ then

$$\hat{e}_i^2 = y_i^2 - b^2x_i^2 \qquad (1.18)$$

Taking summation for i= 1,2,...,N we receive

$$\sum y_i^2 = \sum e_i^2 + b^2\sum x_i^2 \qquad (1.19)$$

On the left-hand side we have the sum of the squared deviations from $\bar{Y}$, which we may take as a measure of the total variation in Y which is to be explained. On the right-hand side, the term $\sum e_i^2$ measures the variation of residuals, i.e. differences between fitted function and observations Y_i and may be taken as variation in Y which remains unexplained by the estimated relationship between X and Y. It follows then that the term $b^2\sum x_i^2$ may be taken as a measure of the variation

in the Y which is explained by the fitted line.

Given the partitioning of total variance into explained and unexplained component we may define a measure of goodness of fit as

$$R^2 = \frac{\text{variation explained}}{\text{variation to be explained}}$$

$$= \frac{b^2 \sum x_i^2}{\sum y_i^2} \qquad (1.20)$$

$$= \frac{\sum y_i^2 - \sum \hat{e}_i^2}{\sum y_i^2} = 1 - \frac{\sum \hat{e}_i^2}{\sum y_i^2} \qquad (1.21)$$

The statistic R^2 is called the coefficient of determination and when it is expressed in the form given in (1.21) it is easy to see that its limits are zero and unity. For example if the fit is perfect, $\sum \hat{e}_i^2 = 0$ and $R^2 = 1$. On the other extreme if all variability of Y is included into residuals $R^2 = 0$. Thus $0 \leq R^2 \leq 1$.

1.7. THE RELATIONSHIP BETWEEN CORRELATION AND REGRESSION ANALYSIS

Because formula (1.8) for b parameter estimator may be presented as

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \qquad (1.22)$$

and formula for coefficient of correlation between two variables can be shown as

then

$$b = r\sqrt{\sum y_i^2} \Big/ \sqrt{\sum x_i^2} = rS_y/S_X \qquad (1.24)$$

A more interesting result emerges if we consider the coefficient of determination R^2. Here if we substitute for b from (1.22) we have

$$R^2 = b^2 \sum x_i^2 \Big/ \sum y_i^2 = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 \cdot \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \frac{\left(\sum x_i\, y_i\right)^2}{\left(\sqrt{\sum x_i^2 \sum y_i^2}\right)^2} = r^2 \qquad (1.25)$$

Thus the coefficient of determination equals the square of the coefficient of correlation between X and Y.

Finally it is worth showing theoretical differences between the correlation and regression models.

1) The correlation model does not specify the direction of a causal link, while the regression model distingushes between the dependent and independent variables.
2) Statistical tests on the correlation model depend upon both X and Y being normally distributed variables, whereas in regression model the X are assumed to be a set of constants and tests are based on the assumption that the random errors are normally distributed.

At the end we would like to show calculation of the standard error S_e and coefficient of determination R^2 for the data on use of fertilizers and yield of wheat.

Having estimated regression equation

$$\hat{Y}_i = 5.5 + 1.14X_i$$

we should calculate residuals

$$\hat{e}_i = Y_i - \hat{Y}$$

and sum of their squares and use them for calculation of S_e. For calculation of R^2 we need additionally $\sum (Y_i - \bar{Y})^2$. All calculations are shown in working Table below.

Yield of wheat (Y_i)-data	Yield of wheat ($\hat{Y}_i$)	Residuals $Y_i - \hat{Y}_i = \hat{e}_i$	Squared residuals $\hat{e}_i^2$	Squared differences $(Y_i - \bar{Y})^2$
8.7	16.9	- 8.2	67.2	56.3
16.0	22.6	- 6.6	43.6	0.0
10.1	11.2	- 1.1	1.2	37.2
7.6	8.9	- 1.3	1.7	74.0
17.0	18.0	- 1.0	1.0	0.6
18.5	16.9	1 .6	2.6	5.3
26.0	21.5	4.5	20.3	96.0
29.0	28.3	0.7	0.5	163.8
27.0	23.7	3.3	10.9	116.6
15.5	12.3	3.2	10.2	0.5
7.5	7.8	- 0.3	0.9	75.7
8.4	5.5	2.9	8.4	60.8
6.2	6.6	- 0.4	1.6	100.0
19.0	14.6	4.4	19.4	7.8
13.0	13.5	- 0.5	0.3	10.2
8.9	10.1	- 1.2	1.4	53.3
15.0	19.2	- 4.2	17.6	1.4
18.0	15.8	2.2	4.8	3.2
31.0	29.4	0.6	0.4	219.0
22.0	20.3	1.7	2.9	33.6
324.4	323.1	1.3	216.9	1115.3

$\bar{Y} = 16.2$

According to formula (1.2) sample variance of dependent variable Y is as follows

$$S_y^2 = 1115.3/19 = 58.70$$

Sample variance of error component

$$S_e^2 = 226.9/(20-2) = 12.6$$

Standard error of estimation

$$S_e = \sqrt{12.6} = 3.6$$

Coefficient of determination

$$R^2 = 1 - \frac{216.9}{1115.3} = 0.806$$

$$\text{or } r^2 = 0.898^2 = 0.806$$

Average dispersion of the data around adjusted straight line is 3.6 . About 81% of the total variance in the yield of wheat is explained by variability in use of fertilizers.

1.8. BIBLIOGRAPHICAL NOTE

The purpose of Chapter I was to present the basic information on the correlation and regression methods and on methods of estimation which are necessary for the topics explored in Chapters II and III.

Here we have confined ourselves to estimation linear association measures between two variables. Some ideas of the nonlinear relations are presented in chapter II. Example of multiple regression model as well as bibliography of the subject are presented in Chapter III.

There is extensive literature on analysis of correlation and regression and on problems of estimation. We may discern here two different approaches to the subject.

Mathematical statistics approach investigates statistical relationships between variables on the base of samples from statistical population. It is exploring statistical properties of different estimators and estimation methods. Representatives of this approach are

among others Draper and Smith [1], Graybill [3], and Hoel [4].

Econometric approach is directed toward construction of models explaining and predicting behaviour of economic variables and is looking for the estimators and estimation methods useful for investigated topics.

Regression models and estimation problems for economic investigations are presented among others in books of Johnston [ ], Klein [6], and Malinvaud [7].

Applications Books for desk programmable calculators HP-29 [11] and HP-67/97 [12] are included to the References as they provide programs for single regression equations.

REFERENCES

1. Draper N.R. and Smith. H. (1966). Applied Regression Analysis. New York: John Wiley.

2. Enslein, K., Ralston, A. and Wilf, H.S. (eds.) (1976) Statistical Methods for Digital Computers. New York: Wiley.

3. Graybill, F.A. (1961). An Introduction to Linear Statistical Models. New York: McGrow-Hill.

4. Hoel, P.G.(1976). Elemetary Statistics. New York.

5. Johnston, J. (1960). Econometric Methods. New York: McGraw-Hill.

6. Klein, L.R. (1956). A Textbook of Econometrics Evanston, Ill.

7. Malinvaud, E. (1970). Statistical Methods of Econometric. Amsterdam: North-Holland.

8. Searle, S.R.(1971). Linear Models. New York: John Wiley.

9. Seber, G.A.F.(1977). Linear Regression Analysis. New York: Wiley

10. Thomas, J.J.(1973), An Introduction to Statistical Analysis for Economists. London: Weidenfeld and Nicolson.

11. Hp-19 C/HP-29C Applications Book. Hewlett-Packard Company 1977.

12. HP-67/HP-97. Statistical Package. Hewlett-Packard Company 1976.

Chapter II

TIME SERIES DECOMPOSITION

2.1. INTRODUCTION

According to data collection method we distinguish data describing some statistical population in one chosen moment or period of time and data on the statistical population in many consecutive equidistanced moments or periods of time. The first kind of data is named cross-section data, the second kind is called time series data.

Below are examples of time series at different levels of agregation:

1. Sequence of consecutive monthly reports of the same family about monthly total expenditures.
2. Sequence of consecutive monthly reports of a factory about production.
3. Esport of coffee from Ethiopia in 1965-1971 E.C.
4. World Production of Wheat in years 1960-1971.

It should be noticed that in some of the above examples the time intervals for observations of time series are one month in others one year. There are time series with even shorter time intervals such as one week or one day and with even longer time intervals such as two or five years.

Time series with daily, weekly, monthly and quarterly observations are short-term observations, time series, with yearly or longer time intervals are long-term observationstime series.

This classification is important because the amount of information about behaviour of examined phenomena in time depends on it how detailed the time series is. Time series with short-term observations contains bigger amount of statistical information than time series with long-term observations. Which kind of time series should be

used for analysis depends on the purpose of analysis. If we are interested only in the long-term tendency we should use yearly data. If one wants to reveal and estimate within year regulatrities should utilize a time series with monthly observations. Ofcourse more labor consuming is short-term time series analysis than long-term one.

We distinguish time series with time interval and momentary(time point) observations. The momentary observation is observation in the determined moment of time, for example temperature at 12^{o} every day. The time interval observation is sum or average of observed statistical variables in the time interval, for example amount of rainfall in a month, average temperature in a month, yearly cars production etc. In principle this last classification of time series is not very important, because these two kinds of time series do not need different methods of analysis.

The most important problems which can be solved by time series methods are the following:

1. Short-term forecasting of the process, that is estimation of the most probable state in the future time T+1, applying forecasting methods which utilize time series observations from the present time T and previous times T-1, T-2,...
2. Decomposition of time series that is "breaking up" the series into trend, cyclical, seasonal and irregular (white noise) components.
3. Control and regulation of the process, consisting in comparing of its actual state with programmed or planned trajectory and in correcting of deviations from it.

These problems are common for different sciences and are solved in different fields or real life such as

economics, sociology, telecommunication, engineering, military, aeronantics etc. The objective of this work are economical application of time series analysis. Particularly, we are interested in decomposition and forecasting problems and methods applied to agricultural statistics.

Decomposition methods are particularly available for analysis of time series with short term observations, such as monthly reports about retail and wholesale prices of foodstuffs, price index of food, monthly reports on food arrivals, monthly or quarterly data about delivery of fertilizers, supply electric power for agriculture etc.

Information about motion of prices may be contaminated with big intra-year and irregular variability. Therefore, to reveal actual tendency of price one should estimate and remove these contaminations.

For some variables, such as demand for fertilizers, farm implements and electric energy the distribution of demand over months for a year is important. Knowing intra year variability of demand it is possible to prepare plan of supply, transportation etc.

Big attention in the work was paid to the simplest time series decomposition technique, i.e. fitting a trend curve to empirical data. However, this technique is very useful for description, in terms of simple mathematical function, the rule of time series development. Apart from that trend adjustment usually is a part of a more complex time series decomposition methods and can serve as a base for a simple forecasting, particularly on the bases of time series with annual observations.

2.2. BASIC DEFINITIONS OF TIME SERIES ANALYSIS

After descriptive introduction to time series analysis we would like to give more strict statistical definitions.

Definition 1

Time series is a sequence of observations which are ordered in time, say x_1 , x_2, ... x_t, ..., x_T, the interval between every two consecutive observations being fixed and constant.

This definition is classical one. More up-to-date and more convenient definition is as follows (compare [14])

Definitions 2

Time series x_1, x_2, ..., x_t, ..., x_T is the sample from the sequence of random variables X_1, X_2,..., X_T. This sample (that is time series) is formed in such way that each random variable is represented by one observation of a time series. Sequence of random variables X_1,..., X_T is called a random or stochastic process.

If we will reflect upon this definition we see that it is not so far from reality. Let us notice that each observation of a time series is the realization of infinite number of potential possibilities. In time t, this infinite number of possibilities is represented by random variable X_t.

For example production of Teff in Ethiopia in 1978/79 (1971E.C.) resulted from thousands of reasons and conditions which caused this production to be about 10 million quintals. Essential for understanding stochastic character of Teff production is that before harvest there were infinite number of possibilities and only one realization of this possibilities occurred (we may say it was sampled by the nature) after harvest. This realization is a time series observation. This stochastic character of time series is the reason that we can not predict an observation of time series exactly. Thus, the theory of probability and statistics can be applied when dealing with time series analysis.

Each random variable X_t and in consequence each observation of time series consists of the two parts: a real or rather regular and white noise which is a random upredictable part of the process.

Thus

$$X_t = X_t^1 + e_t \qquad (2.1)$$

The distribution of e_t is usually assumed to be normal with

$$E(e_t) = 0 \qquad (2.2)$$

and

$$E(e_t\ e_{t+i}) = \begin{cases} \sigma^2 & \text{if } i=0 \\ 0 & \text{if } i\neq 0 \end{cases} \qquad (2.3)$$

How to estimate the regular part of the process (2.1) and its irregular part or in other words white noise, having only one observation from each random variable constituting the random process? Solution of this problem is one of the main tasks of time series methods.

Time series with short-term observations (weekly, monthly or quarterly) consisting of sufficient number of years, contain information about development of this phenomena in the long period of time and about regular variations within yearly periods. First kind of information that is information about direction and speed of changes in long period is called trend, second kind of information, that is information about intra year regular variations is called seasonal, or more general, periodical variations. This two kinds of variations constitute regular part of time series (or more general of the random process) variability. This regular changes are disturbed by irregular (random) variations, which in the theory of random (stochastic) processes is called white noise.

We are going to use for it traditional terms: irregular term or irregular variations.

Time series with long-term observations (yearly or longer)

contains only information about trend-cycle and irregular variations from year to year.

2.3. METHODS OF TIME SERIES PRESENTATION

Time series can be presented in the numerical or graphical form. Numerical presentation is very simple. Single time series usual is presented in the form of sequence of data.

Set of time series should be presented in the form of table. For example production of Teff, Yield of Teff and Area under Teff in 1967-1971 E.C. can be presented as in Table 1.

Table 1

PRODUCTION, YIELD OF TEFF AND AREA UNDER TEFF IN 1967 - 1971 E.C.

Years	Production (in th.qu)	Yield (in qt)	Area (th. hectares)
1967	8264	6.9	1189
1968	9730	6.9	1403
1969	9812	7.5	1311
1970	10009	7.8	1279
1971	10601	7.8	1365

Such arrangement of interrelated data gives possibility to make first visual evaluation of this data.

Time series with monthly or quarterly observations may be presented in the form of table with months arranged as rows or as columns. Example of such method of presentation is given in Table 2. Such arrangement of data enables to reveal intra-year variation and trend.

Table 2

ELECTRICITY SALES FOR AGRICULTURE
IN 1973 - 1977

Years	Quarters			
	I	II	III	IV
1973	134	115	91	144
1974	154	265	97	299
1975	394	191	86	150
1976	168	140	95	219
1977	217	186	151	195

Sometime more informative about over-time behaviour of time series may prove graphical presentation. Example of graphical presentation is given in Fig. 1

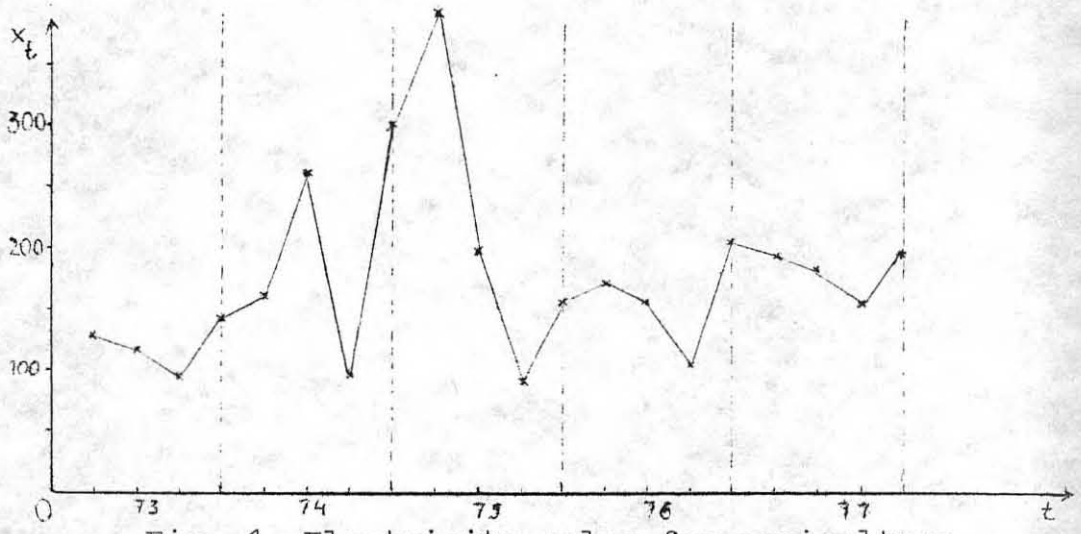


Fig. 1. Electricity sales for agriculture in 1973-1977

The horizontal line known as "t axis" represent time: years and months within years. The vertical line is destinated for the observations of time series. The point at which the two axes intersect is zero for both of them and is referred as zero point or origin. For monthly or quarterly data it is necessary to divide time scale for years

and months (quarters). From the technical point of view it is important to choose proper proportion for a curve diagram. If time axis is exaggerated diagram can give an impression of unimportant fluctuations of examined phenomena. When vertical scale is exaggerated the diagram can give an impression of tremendous fluctuations. Problem of proper scale choice is very important when we want to compare an over-time behaviour of a few related time series. It is recommended to use the same scale for all sets of data presented in the same units. For example if we want to compare area under different cereals in Ethiopia, we may use the same vertical and horizontal scales for all diagrams or to present them within common t and X axis. But sometime the magnitudes of compared data are so heterogeneous that similar scale do not give possibility to assess if rates of their changes are similar or not. Such situation is presented in Fig.2. where following sequences of numbers were plotted

t	1	2	3	4	5
X1	5	6	7.2	8.6	10.4
X2	30	36	43.2	51.8	62.2

each of the same rate of growth equal 20%. In Fig.3 the same numbers after logarithmic transformation were plotted.

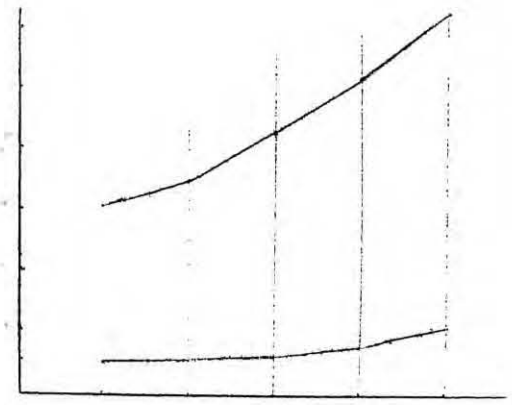


Fig.2. Plot of original data

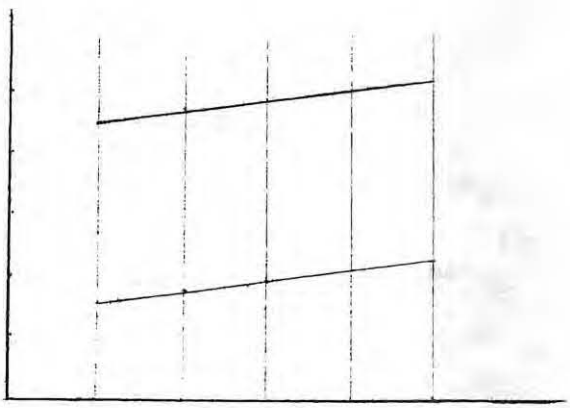


Fig.3. Plot of logarithmically transformed data from Fig.2

The ability of the usual type of chart to give a satisfactory visual impression of absolute change but not of ratio of change is brought out by the example below which we repeat after Croxton and Cowden [5] .

Table 3

An arithmetic (A), geometric (G) progression and progression with non constant ratio of increase (NC)

Year (Xvalue)	(A) Arithmetic		(G) Geometric		NC		
	Y value	Amount of increase	Y value	Percent of increase	Y value	Amount of increase	Percent of increase
1971	0		128		50		
1972	200	200	192	50	80	30	60.0
1973	400	200	288	50	160	80	100.0
1974	600	200	432	50	300	140	87.5
1975	800	200	648	50	550	250	83.3
1976	1000	200	972	50	1080	530	96.4
1977	1200	200	1458	50	1730	650	60.2
1978	1400	200	2187	50	2500	770	44.5

In Table 3 three kinds of changes are presented. The case A represents a constant amount of increase of 200 units per year. This or any other arithimetic progression (constant amount of increase or decrease) will be depicted by a straight line when plotted on the arithmetic grid. This case is shown in Fig. 4 by straight line A. Curve G is the result of plotting a series of figures for the G set of data from Table 3. Observations of this time series show increase of 50% each year. This curve bends upward more and more sharply as time passes.

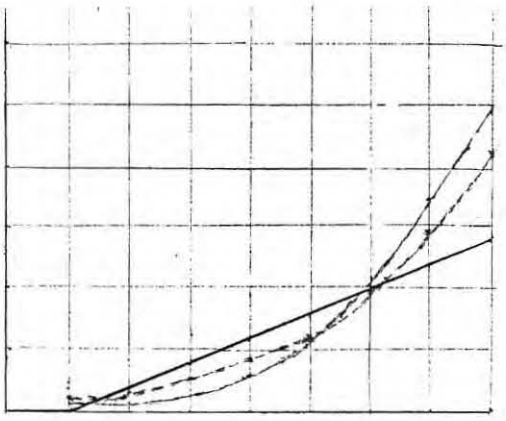


Fig.4 Arithmetic progression (A), geometric progression (G) and curve with non-constant ratio of increase (NC).

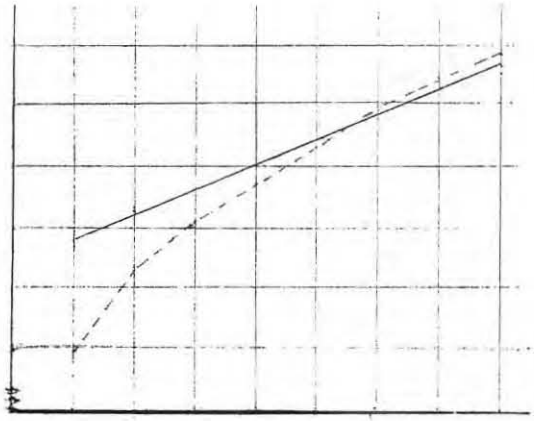


Fig.5 Geometric progression (G) and progression with non-constant ratio of increase NC plotted on semi-logarithmic grid.

A series showing a constant ratio of increase or decrease is known as a geometric progression, and any geometric progression will yield a curved line when plotted on arithmetic grid. An increasing geometric progression (trend) is represented by a curve line which slopes upward and is concave upward; a decreasing geometric progression is represented by a curve which slopes downward and is concave upward. A serious difficulty in interpreting such curves, however, lies in the fact that the eye cannot discern whether or not a particular curved line does or does not represent a constant ratio of change. Curve NC depicts a NC set of data from Table 3 which is neither an arithmetic progression nor geometric. An eye cannot notice this fact because the curve bends upward. Therefore it is not possible for the reader of an arithmetic chart to be sure whether a curved line represents a constant ratio of increase, a ratio of increase which is diminishing or a ratio of increase which is accelerating. Any series of figures that increase more rapidly than an arithmetic progression (for example 10, 12, 15, 24, 30) slopes upward and is concave upward when is plotted

on arithmetic grid; any series of figures that decreases less rapidly than an arithmetic progression (for example 100, 91, 83, 76, 70, 65) slopes downward and is concave upward when shown on arithmetic coordinates.

Therefore we may conclude that graphic comparisons in respect to ratios of changes will be facilitated if we can employ a kind of grid which will make a constant ratio of increase or decrease appeared as a straight line.

It can be done in two ways. We can make a diagram of ratios of X values instead of their absolute (arithmetic) values. Such diagram has an excellent diagnostic property because constant rate of change (increase or decrease) gives a straight line parallel to a t-axis, linearly accelerating ratio of increase gives a straight line sloped up, linearly diminishing ratio of increase given a straight line sloped down, etc. Similar result we will receive if for diagnosis purpose we use instead of ratios logarithms of X_t values and draw them against the background of t values, because logarithms of constant geometrically progressive numbers have constant first differences and therefore their diagram gives a straight line.

Table 4

A geometric progression and logarithms of the geometric progression

year (X value)	y value	Logarithm of y value	Amount of increase of logarithm
1971	128	2.10721	
1972	192	2.28330	0.17609
1973	288	2.45939	0.17609
1974	432	2.63548	0.17609
1975	648	2.81157	0.17609
1976	972	2.98766	0.17609
1977	1458	3.16375	0.17609
1978	2187	3.33984	0.17609

In Table 4 the geometrically progressed data from Table 3 are shown again and with it are given the logarithms of the numbers and their first differences (signed as amount of increase of logarithms). Examination of this logarithms reveals that they form an arithmetic progression; therefore, if this logarithms are plotted on an arithmetic grid a straight line will result, as may be seen in Fig. 5(line G). Method consisting in plotting of logarithms instead of original values of X_t series involves the additional step of looking up logarithms before the data can be plotted.

However, instead of plotting the logarithms of the observations we may use a grid which is designed with a logarithmic vertical scale as in Fig.6. Semilogarithmic grid can be bought or can be prepared. It can easily be done because it merely involves spacing the vertical scale values in proportion to the differencies of their logarithms. For example the semilogarithmic scale presented in Fig.6 was prepared by spacing vertical scale in proportions between logarithms of consecutive natural numbers from 1 to 64. This sequence of numbers may be replaced by any other for example 15, 30, 45..., or 0.5, 1.0, 1.5..., because semilogarithmic scale is proportional to the differences of logarithms of numbers, and this differences depend only on the ratio of numbers not their magnitudes. Therefore, the same difference of logarithms stands between 100 and 200 and 0.5 and 1.0 because this two pairs of numbers has the same ratio 1:2. Great advantage of the semilogarithmic scale is that we can compare ratios of changes of phenomenas of different scales (magnitudes) as it was shown earlier in Fig.2 and 3.

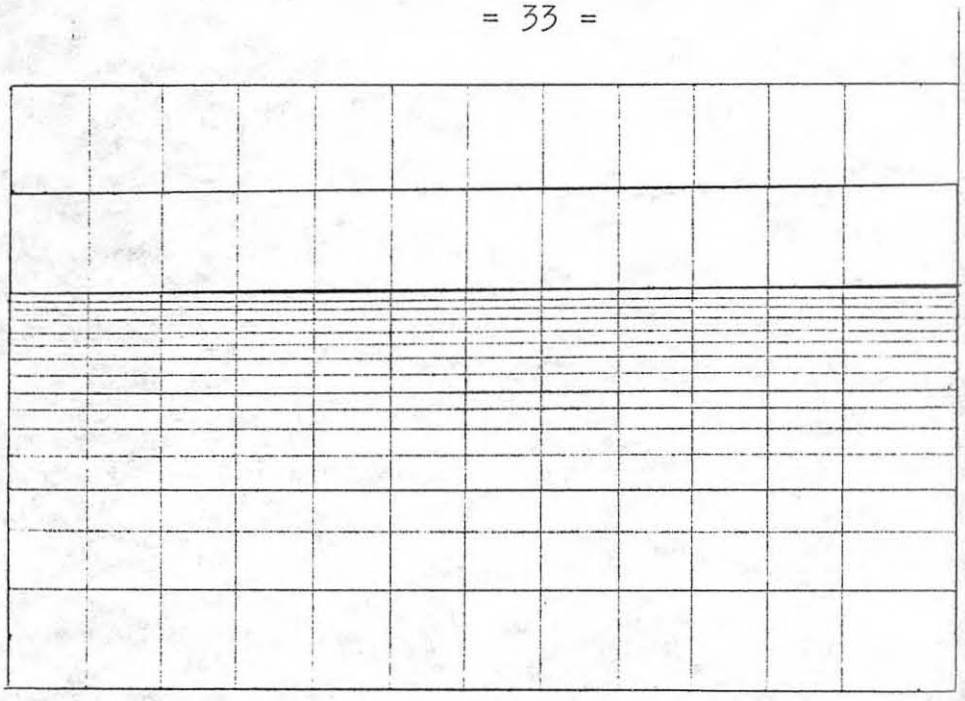


Fig.6. Semilogarithmic scale for numbers with ratios as 1:2:3:...:64.

2.4. METHODS OF TREND ESTIMATION

2.4.1. General problems of trend estimation

There are three important reasons for attempting to estimate the trend. The first reason is to investigate the deviations from the trend. These deviations (residuals) can consist of seasonal and irregular variations. Sometimes when time series includes sufficiently long period of time, this time series of deviations includes information about cyclical variations. If time series is not very long, covering only part of a cycle it is analysed as part of the trend.

Second reason is that we want to study the trend itself in order to compare one trend with other, to analyse changes of trend in connection with different

factors influencing it, and to forecast future value of trend.

These reasons cause that in many methods of t.s. decomposition estimation of a trend is a first stage of decomposition procedure.

The purpose for which measurements are made partly determines the methods adopted. If the object is solely to isolate trend and cycles it seems reasonable to suppose that the choosen trend line should pass through the cycles in such way as approximately to allow a balancing between the positive or negative phases of each cycle. When our task is not isolating trend and cycle but only to follow the cycle, trend and cyclical variations should be considered jointly. When the object is to make comparisons, generalizations or forecasts, the curve should be expressed by a mathematical formula. By means of such a formula one can, for instance say that at a given time series shows a certain ratio or certain amount of growth per annum or per month and that if this tendency continues, the trend will reach a certain value at some specified time in the future. Fitting a trend by a mathematical formula does not, however, remove the subjective element from trend fitting. The statistician can vary the behaviour of the curve by selection of the type of formula he employs or the years to which he fits the curve.

A mathematical equation allows us not only to draw the trend of a time series but also provides in the trend equations, a consise definition of that trend. If the trend itself to be studied or to be extended is beyond the observed data it is particularly desirable to describe it by an objectively determined equation.

2.4.2. Fitting a straight line

The simplest type of curve is the straight line, which is described by an equation of the type.

$$p_t = a + bt \qquad (2.4)$$

in which p_t are the points of the straight line coresponding to the values of time t. Variable t is called time variable and denotes nominal units of time, for example years: 1967, 1968,... or, more often, numbered units of time for example 0, 1, 2,..., T where numbers are given as numbers of consecutive observations of time series. Because straight line (2.4) is adjusted to time series data, x_t, we should rewrite this equation using symbols of time series observations, but it should be kept in mind that because of irregular and periodical variations straight line do not pass through the time series observations. Differences between true observations and trend will be called residuals.

Therefore, relations between t.s. observations and trend line p_t should be presented as follos:

$$x_t = p_t + r_t \qquad (2.5)$$

where r_t - residuals.

For time series with yearly observations this residuals are simple irregular variation which according to our convention we denote with e_t. Then for such time series for which straight line is true, trend line can be presented as follows:

$$x_t = a + bt + e_t \qquad (2.6)$$

parameters a and b are parameters of <u>unknown</u> trend line of the random process.

Parameters of trend line adjusted to the sequence of t.s. observations we will denote as $\hat{a}$ and $\hat{b}$ and trend value calculated with this parameters as $\hat{x}_t$. We use symbols $\hat{a}$ and $\hat{b}$ because estimated parameters are only approximation to the true parameters. Therefore, approximation of the unknown true trend is trend line described by the following equation:

$$\hat{x}_t = \hat{a} + \hat{b}t \qquad (2.7)$$

Values a, b and x_t depends on many things such as a length of t.s. taken to the analysis, applied estimation procedure etc. Straight line can be adjusted to the set of data presented on the diagram by hand. It is the simplest estimation procedure.

For t = 0 equation (2.7) becomes

$$\hat{x}_o = \hat{a} \qquad (2.8)$$

Then parameter a is the initial value of linear trend (when t=0). It is a fixed point at which trend line intersect x_t - axis. Parameter b defines relation between t variable and trend line x_t. This relation is such that absolute trend value is proportional to the corresponding value of t plus parameter a, that means that b can be calculated (from the diagram of trend line adjusted by hand) as follows:

$$b = \frac{x_t - a}{t} \qquad (2.9)$$

Then parameter b determines the "speed" of a straight lines change.

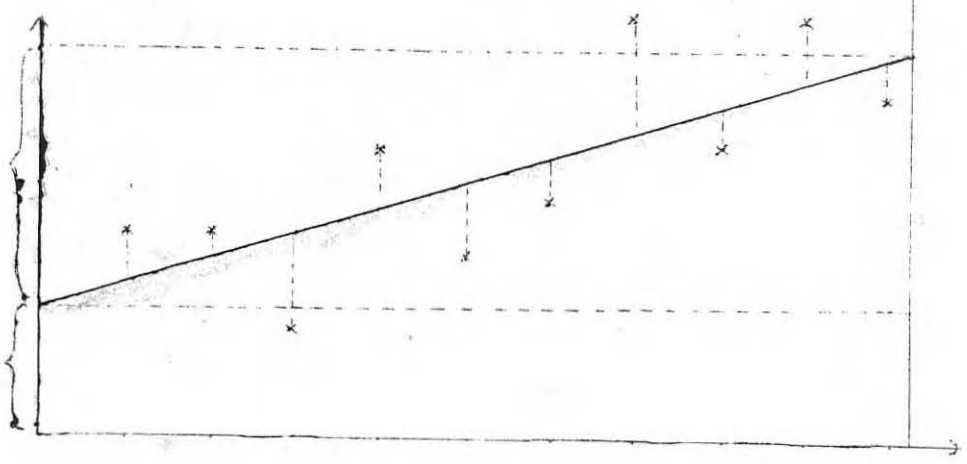


Fig.7. A straight line adjusted by hand to the dispersion of time series observation.

Value of b is the same for any t, for example (Fig.7):

$$t = 10, \hat{x}_{10} = 3, \hat{a} = 1, \text{ then } \hat{b} = (3-1)/10 = 0.2.$$

Parameter b defines the slope of the straight line or in other words the angle between trend line and t - axes. If $\hat{b} = 1$ it means that $\hat{x}_t = \hat{a} + t$. If $\hat{b} = 0$, $\hat{x}_t = \hat{a}$ (constant).

2.4.3. A least squares method

The straight line can be adjusted to the time series more exactly using analytical method. Adjusting of a straight line means merely estimation of a and b parameters.

Simple and exact method of the parameters estimation in the case when relation between variables is linear is least squares method.

The least squares criterion applied to the fitting of a straight line trend to time series can be expressed as follows (see chapter I, section 1.3):

$$D = \sum e_t^2 = \sum /x_t - \hat{x}_t/^2 = \text{minimum} \qquad /2.10/$$

where $\hat{x}_t$ is expressed by /2.7/.
Taking into consideration that $\hat{x}_t$ stands for a straight line, the formula /2.10/ can be rewritten to the form:

$$D = \sum /x_t - a - bt/^2 = \text{minimum} \qquad /2.11/$$

Parameters a and b can be found from the following equations:

$$\frac{\partial D}{\partial a} = 0 \quad \text{and} \quad \frac{\partial D}{\partial b} = 0 \qquad /2.12/$$

where $\frac{\partial D}{\partial a}$ and $\frac{\partial D}{\partial b}$ are partial derivates of function D with respect to a and b . This gives the set of so called normal equations /see chapter I, section 1.3, formula 1.7 /.

Solving this set of equations with respect to a and b we find

$$b = \frac{\sum x_t t - \frac{\sum x_t \sum t}{N}}{\sum t^2 - \frac{(\sum t)^2}{N}} \qquad /2.13/$$

$$a = \bar{x} - b\,\bar{t} \qquad /2.14/$$

where $\bar{x}$ is the arithmetic average of x_t,

$\bar{t}$ - the arithmetic average of t .

We can transform t variable in such way that formula /2.13/ will be simplified. When instead of t = 1, 2, 3,..., T, we use variable $t' = t - \bar{t}$, parameter b is calculated as follows:

$$b = \frac{\sum x_t t'}{\sum t'^2} \qquad /2.15/$$

because $\sum t' = o$, and $/\sum t' /^2 = o$.

Formula (2.15) gives the same estimation of parameter b as estimation on the base of formula (2.13), but is simpler. Below are given examples of the t variable transformed.

1) t= 1,2,3,4,5 (T= odd number)
   $\bar{t}$= 3 t' = -2, -1, 0, 1, 2

2) t= 1,2,3,4,5,6 (T= even number)
   $\bar{t}$= 3.5 t' = -2.5, -1.5, -0.5, 0.5, 1.5, 2.5

3) t= 1967, 1968, 1969, 1970, 1971
   $\bar{t}$= 1969 t'= -2, -1, 0, 1, 2.

Exercise 1.

Using least squares method adjust a straight line trend to the data on the Area under Teff in Ethiopia in 1967 - 1971E.C.

Working Table

D A T A			t'^2	$x_t t'$	x_t (calculated trend observations)	Residuals $x_t - x_t = e_t$
x_t	t	t'				
1189	1	-2	4	-2378	1263	-74
1403	2	-1	1	-1403	1286	117
1311	3	0	0	0	1309	2
1279	4	1	1	1279	1332	-53
1365	5	2	4	2730	1354	11
$\sum x_t$= 6547 $\bar{x}$=1309			$\sum t'^2$= 10	$x_t t'$= 288	$\sum x_t$ = 6544	3

$$\hat{b} = \frac{228}{10} = 22.8$$

$$\hat{a} = 1309 - 22.8 \cdot 3 = 1240.6$$

Equation of a straight line:

$$\hat{x}_t = 1240.6 + 22.8$$

Average increase of area under teff per annum in 1967-1971 E.C. was 22.8 th. hectares.

2.44. GOODNESS OF FIT OF ADJUSTED TREND

Goodness of fit of adjusted trend is measured by means of the standard error of adjustment

$$S_e = \sqrt{\sum (x_t - \hat{x}_t)^2/(T-2)} \qquad (2.16)$$

which shows average dispersion of the time series observations around the trend line or in other words average dispersion of residuals.

The number by which sum of squares is divided in (2.16) formula, T-2 is called a number of degrees of freedom. The reason for which we should subtract from T just 2 needs a complex mathematical explanation. Here we explain only that degrees of freedom are equal to the number of observations minus number of estimated parameters in fitted model.

Coefficient of determination R^2 is the relative measure determining what part of the total time series variance is explained by a fitted trend or more general by a model fitted to time series (compare Chapter I, section 1.6)

2.4.5. EXPONENTIAL TREND

In the t.s. analysis extensive use is made of trends showing relative variations of the consecutive trend values as constant. It means that ratio of the every two consecutive trend values is the same: $x_t/x_{t-1} = \beta$. If $\beta > 1$ it means that trend expressed in the absolute value is increasing. For decreasing trend should be $\beta < 1$.

Analytical trend with constant relative variations adjusted to the empirical time series looks like follows:

$$x_t = \alpha\beta^t \, e_t \qquad (2.23)$$

where α and β are parameters, e_t residual term. It has to be kept in mind that residual term e_t for the time series with annual observations includes irregular variations, for time series with monthly or quarterly observations it can include both periodical and irregular variations.

The diagram of it we can present as a trend line passing throught the dispersion of t.s. observations:

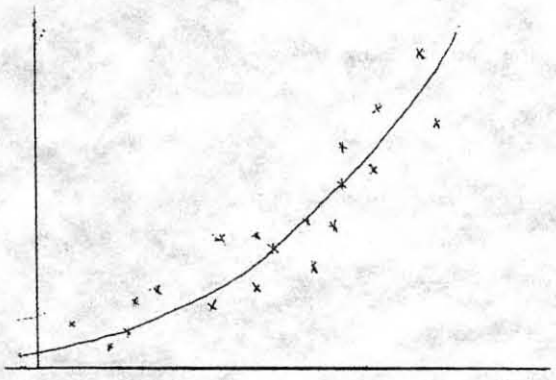


Fig.8. Increasing geometric progression, $\beta>1$

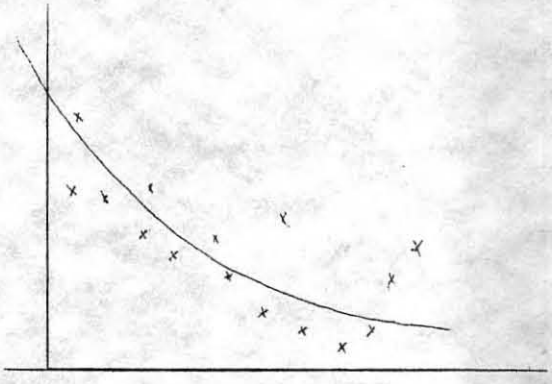


Fig.9. Decreasing geometric progression, $\beta<1$

It follows from the mathematical rule that logarithms of the number composing a geometric progression, form the arithmetic progression, diagram of which is a straight line. This rule shows that adjustment of a trend line to the time series whose observations tend to form a geometrical progression can be easily performed after logarithmic transformation of time series observations. After such transformation we only need to adjust a straight line to the logarithms.

After logarithmic transformation equation (2.23) becomes an equation of straight line

$$\log x_t = \log \alpha + t \log \beta + \log e_t \qquad (2.24)$$

when we substitute $\log \alpha = a$, $\log \beta = b$ and $\log e_t = e'_t$ we receive

$$\log x_t = a + bt + e'_t \qquad (2.25)$$

The straight line adjusted to logarithms of observations can be presented as follows:

$$\log \hat{x}_t = \hat{a} + \hat{b}t \qquad (2.26)$$

From the above equation follows that original geometric progressive trend x_t is the set of antylogarithms of the straight line points calculated on the base of (2.26)

Parameters a and b can be estimated by means of the least squares method. It means that the method of least squares will now satisfy condition that the sum of squared differences between the logarithms of the time series observations and logarithms of the corresponding trend values shall be a minimum. Parameters a and b are calculated using the same formulas as for a straight line, only instead of original x_t values their logarithms are employed.

In equation (2.23) constant β can be presented as $(1+q)$ and α can be substituted by x_0. Then equation can be presented as follows:

$$x_t = \hat{x}_0 (1 + q)^t e_t \qquad (2.27)$$

Parameter $\hat{x}_0$ is the value of trend for t= 0. Parameter $\beta = (1 + q)$ is called a coefficient of growth, is the rate of growth. Equation of estimated trend line is

$$x_t = x_0 (1 + q)^t \qquad (2.28)$$

Equation (2.27) is important in statistics, partcularly in economical statistics because geometric progression or in other words trend with a constant rate of growth very often gives a good approximation to the development of the real economical and social processes. It should be mentioned that the usual practice in the planning, designing, formulation of development programs etc., is to present anticipated changes as a percentage of the state from the past.

2.4.6 Average rate of growth

Rate of growth is sometimes applied as a measure of average growth when we want to characterise shortly the development of examined phenomena from time 0 to time T. According to formula (2.28) amount in time 0 is x_o in time T is x_T. Average rate of growth between T and 0 units of time may be calculated on the bases of the transformed formula (2.28), i.e.

$$(1+q)^T = x_T/x_o$$

$$(1+q) = \sqrt[T]{x_T/x_o} \qquad (2.29)$$

$$q = \sqrt[T]{x_T/x_o} - 1$$

Formula (2.29) is geometric average of relative number

$$x_T/x_o$$

Exercise 2

We want to calculate average rate of growth of the value of coffee exported from Ethiopia from 1961 to 1976. Value of exported coffee in 1961 was x_o = 93.6 mln. Eth. Birr, in 1976 x_T = 324.6mln. Eth. Birr.

$$x_T/x_o = 3.468$$

$$1+q = \sqrt[15]{3.468} = 1.0864$$

$$q = 1.0864 - 1 = 0.0864$$

Coefficient of growth of exported coffee was 108.64%, the rate of growth was 8.64%.

From (2.29) follows that formula for the average rate of growth does not take into consideration values from intermediate years between 0 and T years. Value is adjusted to data in such way that starting from x_0 it exactly reproduces x_T value. For example we can calculate the successive theoretical values of exported coffee, going out from the x_0 = 93.6 and applying the calculated rate of growth = 8.64%, as follows:

$$x_1 = x_0 \cdot 1.0864 = 93.6 \cdot 1.0864 = 101.7$$
$$x_2 = x_1 \cdot 1.0864 = x_0 \cdot 1.0864^2 = 93.6 \cdot 1.0864^2 = 110.4$$
$$x_3 = x_2 \cdot 1.0864 = x_0 \cdot 1.0864^3 = 120.0$$
$$\vdots$$
$$x_{15} = 93.6 \cdot 1.0864^{15} = 93.6 \cdot 3.468 = 324.6$$

Therefore the average rate of growth should be applied and interpreted very carefully particularly when comparing rate of growth of several phenomenas with similar ratios x_T/x_0 but with different trajectories of the development.

Example 1

We want to compare rate of growth of the following two process:

t	0	1	2	3
A	100	200	200	200
B	100	50	50	200

For both of them ratio $x_3/x_0 = 2$ and rate of growth is $q = \sqrt[3]{2} - 1 = 0.26$. As one can see trajectories from the x_0 to x_T values are for A and B quite different, but their evaluation on the base of q is the same.

Then we come to the conclusion that every application of as a measure of the average rate of growth should be preceded by the analysis of intermediate data.

Exercise 3

Adjust exponential curve to data on the area under Vetch in Ethiopia in 1967-1971 E.C., compare it with result of a straight line fitting from Section 4.3 using Se and R^2 measures.

Working Table

Data x_t	t	t'	$\log x_t$	$t' \log x_t$	t'^2	$\hat{x}_t$	$e_t^2 = (x_t - \hat{x}_t)^2$	$x_t - \bar{x}$	$(x_t - \bar{x})^2$
29	1	-2	1.4624	-2.9248	4	29	0	-16	256
32	2	-1	1.5051	-1.5051	1	35	9	-13	169
50	3	0	1.6990	0	0	43	49	5	25
54	4	1	1.7324	1.7324	1	52	4	9	81
58	5	2	1.7634	3.5268	4	63	25	13	169
223			8.1623	0.8293	10		87		700

$\sum t' \log x_t = 0.8293$, $\sum t'^2 = 10$

$\hat{\beta} = 0.8293/10 = 0.0829$

Mean of $\log x_t = 8.1623/5 = 1.6325$

$\hat{\alpha} = \bar{x} - b\bar{t} = 1.6325 - 0.0829 \times 3 =$

$= 1.3838$

Adjusted trend function:

$$\hat{x}_t = 1.3838 + 0.0829t \qquad (2.30)$$

Theoretical (calculated on the base of fitted function) values are following:

x_1 = antilog (1.3838+0.0829)= $10^{1.4667}$=29
x_2 = antilog (1.3838+0.0829·2) = $10^{1.5496}$= 35
x_3 = antilog (1.3838+0.0829·3) =$10^{1.6325}$= 43
x_4 = antilog (1.3838+0.0829·4) = $10^{1.7154}$= 52
x_5 = antilog (1.3838+0.0829·5) = $10^{1.7983}$= 63

$$\sum (x_t - x_t)^2 = 87$$

S_e^2 = 87/3 = 29 S_e = $\sqrt{29}$ = 5.39

Average dispersion of t.s. observations around exponential trend line is 5.39

$$\bar{X} = \frac{1}{T}\sum x_t = 223/5 = 44.6$$

$$\sum (x_t - \bar{x})^2 = 700$$

$S^2(x_t)$ = 700/4 = 175 $S(x_t)$ = $\sqrt{175}$ = 13.2

Average dispersion about mean value is 13.2.

R^2 = 1- 87/700 = 0.876

Conclusion

Exponential trend explains 87.6% of the total t.s. variability measured by means of t.s. variance. We can compare goodnes of fit measures for exponential trend with measures for the straight line adjusted to the same t.s.,i.e.

S_e^2 = 17.3 S_e = 4.1

R^2 = 1 - 17.3 /175 = 1 - 0.099 = 0.901

Then straight line trend includes about 90% of the total t.s. variability measured by means of the t.s., variance. The conclusion is that the straight line trend better fits to the considered time series.

2.4.7. Miscellaneous analytical trend models

Parabolic power and logarithmic trends

Curves of the parabolic (power) and logarithmic type

$$x_t = at^b \qquad (2.31)$$

$$x_t = a + b \log t \qquad (2.32)$$

are not generally applicable to time series, as the time variable can not in strict logic, be raised to any power or logarithmed. However their use may be justified on empirical grounds, if they hapen to describe a given trend accurately. The shape of power and logarithmic curves are shown in Fig.10. and 11.

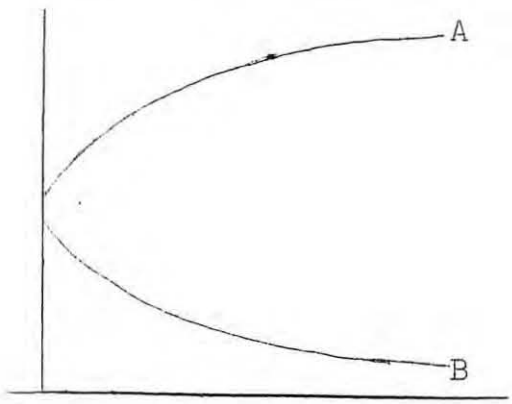


Fig.10. (A) parabola with parameter $b > 0$, (B) parabola with $b < 0$.

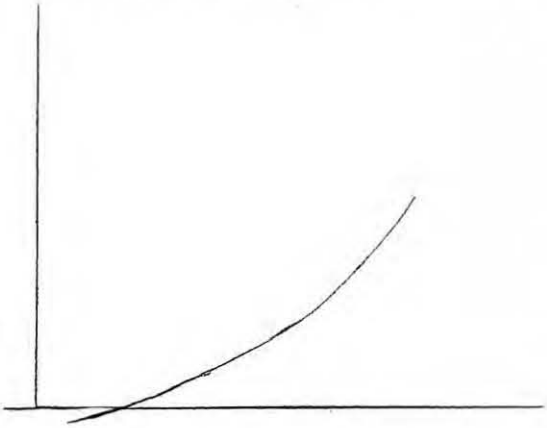


Fig.11. Logarithmic curve

On logarithmic scale parabolic curve becomes a straight line (Fig.12)

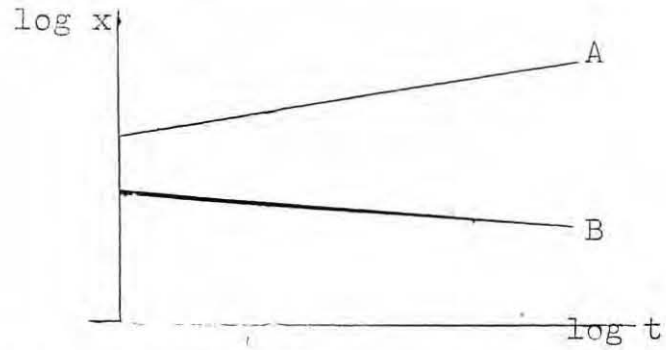


Fig.12. parabolic curves (A) and (B) plotted on the logarithmic scale are transformed to a straight lines.

Parameters of the functions (2.31) and (2.32) may be estimated by least squares method. The power function (2.31) should be transformed to linear by means of logarithmic transformation

$$\log x_t = \log a + b \log t \qquad (2.32)$$

$$\text{or} \quad \ln x_t = \ln a + b \ln t \qquad (2.33)$$

Parameters log a and b can be calculated by means of (2.13) and (2.14) formulas respectively only instead of x_t and t we should use $\log x_t$ and log t.

From (2.31) (2.32) and (2.33) follows that

$$a = 10^{\log} \text{ or } a = e^{\ln}$$

$$x_t = 10^{\log x_t} \text{ or } x_t = e^{\ln x_t}$$

Estimation of the (2.32) function is straight forward. Parameters a and b are calculated on the base (2.13) and (2.14) formulas only log t (lnt) is used in place of variable t.

Gompertz curve

$$x_t = ab^{c^t} \qquad (2.34)$$

has sometimes been considered as expressing the low of growth of a human population. This last curve has two asymptotes (upper and lower).

Asymptote is a value to which a curve tends us it approaches the infinity.

This curve represents the kind of dying or fading growth. It is applied to the approximation of the changes of processes whose development trajectory consists of three stages: 1) initial slow growth, 2) fast increase 3) increase with diminishing rate of growth.

Parameters a, b and c cannot be estimated by means a such straight forward method as least squares. Methods of estimation are presented in [3,21] .

Polynomial trend

The simplest analytical trend is a straight line adjusted to the original or logarithmically transformed data. Such trend approximates the changes with constant increments of absolute values of observations or their logarithms (constant relative changes). A straight line is called polynomial of the first degree.

If a straight line does not fit the data we try a curve of higher degree than first, for example

second degree polynomial (parabola)

$$x_t = a + bt + ct^2 \qquad (2.35)$$

or third degree (cubic)

$$x_t = a + bt + ct^2 + dt^3 \qquad (2.36)$$

For practical purposes no curves beyond the third degree are used, as curves of higher degrees have too many

bends (i.e. changes in direction from positive to negative or vice versa) to be consistent with our definition of trend as a broad and gradual, smooth, longterm movement over the whole period considered, such as persistent growth or persistent decline. Moreover high degree polynomials even though they show the best measures of the goodnes of fit, can give very bad trend extrapolations, because estimated trend is too sensitive to the irregular changes of data.

Second degree polynomial trend should be interpreted as trend with decreasing or increasing its first differences but constant second differences. Third degree polynomial trend show its third differences as constant. Equations (2.35) and (2.36) contain respectively three or four parameters. Least squares formulas for this parameters presented in the usual way are very complicated. For their presentation is applied matrix notation. It is possible to calculate them using more sophisticated desk calculators [26], [27], or special computer programs for polynomial approximation or for regression analysis (compare [28]). Sometimes polynomial approximation is included to the special computer programs for time series decomposition.

Polynomials such as (2.35) and (2.36) can be adjusted to the logarithmic transformations of t.s. observations. It means that we fit exponential curve of the second or third degree to the original data. For example

$$\log x_t = a' + bt + c't^2 + e_t \qquad (2.37)$$

means that

$$x_t = a\, b^t\, c^{t^2}\, e_t \qquad (2.38)$$

where a; b', c and abc are parameters a'= log a, b'=log b

and $c' = \log c$, e_t - irregular term of time series. Equation (2.38) describes exponential growth with constant second ratios of estimated trend $\hat{x}_t$.

It means that

$$q_{1t} = \hat{x}_t / \hat{x}_{t-1} = bc^{2t+1}$$

$$q_{2t} = q_{1t} / q_{1t+1} = c \quad \text{(constant)}$$

i.e. constant second differences of $\log x_t$ show that second ratios of original x_t are constant.

2.4.8. Selection of the curve to represent trend

There is no single general rule for the choice of trend curve. From the definition of the notion of trend it follows that the curve shold be as simple as possible. Sometimes, if there are structural changes during the period considered it is necessary to break up that period into parts, fitting separate trend lines to each part.

If it is impossible to determine the appropriate curve by subjective inspection we may use some objective tests based on the relations between t and x_t. By arranging the natural values of t in an arithmetic series we get a series of corresponding successive values of x_t whose differences are called first differencies

$$\Delta x = x_t - x_{t-1}$$

second differencies

$$\Delta^2 x = \Delta x_t - \Delta x_{t-1}$$

third differencies $\Delta^3 x$, and so on.

If the first differencies are approximately constant the trend is expressed by a straight line, if the second differencies are constant the trend can be expressed by a second degree polynomial and so on.

Similar tests may be applied to differences not between the natural value of x_t but between their logarithms or their reciprocals.

If the first differences of the logarithms are approximately constant the trend of the t.s. will be expressed by an exponential curve (i.e. it will be possible to fit a straight line to the logarithms) if the second differences of the logarithms are constant it will be possible to fit a second degree polynomial to the logarithms and so nn.
If the first differences of the logarithms are declining by approximately constant percentage the trend will be expressed by a Gompertz curve.

2.4.9. The method of moving averages

One alternative to fitting a single high order polynomial to the whole t.s. is to fit lower-order polynomials to sections of t.s. This might be done by dividing the t.s. up into a number of sub-periods to which lower-order polynomials were fitted by least squares. This has the disadvantage that there may be no obvious criterion for splitting t.s. into sub-periods and that there would be problems with the discontinuities in moving from one sub-period to another.

One alternative, which avoids both problems by not specifying separate sub-periods is the method of moving averages.

The simplest application of this technique may be described as follows. Suppose we have a total of T equally spaced observations on time series x_t, and choose a subperiod of length r ($r<T$). We then take the first r observations and calculate the arithmetic mean

$$\frac{1}{r} \sum_{t=1}^{r} x_t \quad .$$

Having computed this value, we move on by dropping the first observation and introducing the (r + 1) th. We then compute

$$\frac{1}{r} \sum_{t=2}^{r+1} x_t$$

Now we move on dropping the second observation and bringing in the (r+2) th observation computing

$$1/r \sum_{t=3}^{r+2} x_t .$$

If the length of our sub-period is an even number the first observation in our new series is located half way between the (r/2+1) and (r/2+2) th observations in the original series and so on. This is only a minor problem, since by taking the average of successive pairs of averages we may centre them on the original observations. If r is an odd number, say r=2m+1, than no problem arises, since the first observation in the new series will be located at the same point in time as the (m+1)th observation in the original series. Let us denote the new series by $x_{t'}$, $t' = (r/2+1), (r/2+2), \ldots, (T-r/2)$.

The averaging process tends to eliminate period to period fluctuations and produces a much smoother series than the original observations. Then we can describe the result

of the averaging as smoothing of t.s. It is to be noted that the method of moving averages has a descriptive rather than an analytical value; it does not determine the rule of variation of the variable x_t as a function of time.

Since the straight line estimated by least squares passes through the mean point $(\bar{x},\bar{t})$ of the dispersion of the data and each moving average value is located in the mean point of each r subset of t.s. observations we can conclude that simple averaging is equivalent to fitting a linear trend by least squares to successive blocks of r observations and only using a medium points of this lines as estimated trend value.

It is possible to fit quadratic or higher-order polynomials instead of straight line. Instead of simple (arithmetic) average of consecutive r observations we use in this last case weighted averages whose weights follow from the higher degree polynomials adjusted to subperiods of r observations. Weighted m.av. so called Spencer's 15- term average can be shown as a set of weights by which consecutive t.s. observations should be multiplied, i.e.

$$\begin{aligned}\bar{x}_{t-7} = {} & -0.009_{x-14} - 0.019x_{t-13} - 0.016x_{t-12} + 0.009x_{t-11} \\ & + 0.066x_{t-10} + 0.144x_{t-9} + 0.209x_{t-8} + 0.231x_{t-7} \\ & + 0.209x_{t-6} + 0.114x_{t-5} + 0.066x_{t-4} + 0.009x_{t-3} \\ & - 0.016x_{t-2} - 0.019x_{t-1} - 0.009x_t\end{aligned}$$

where t= 15, 16, ..., T

Such moving average can reproduce up to cubic (third degree) polynomial trend, but has the disadvantage that don't remove exactly the seasonal variability.

Such moving averages are applied in combination with simple moving averages in iterative methods of time series decomposition (compare section 2.5).

One question which remains to be answered is how the value of r (the length of moving average) is to be chosen. The rule is the longer is the moving average the smoother the resulting series will be, so that moving average should be long enough to eliminate periodical fluctuations in the t.s. However r observations are lost, since there are no values of the smoothed series corresponding to the first r/2 and last r/2 observations in the original series. The rule, which is often applied is that one should use the shortest moving average which enables smoothing of the periodical variation.

Example 2

1) Calculate 4 term moving average of the x_t t.s. and centralize it on t= 3,4,5,6,...,2) Calculate 5 term moving average x_3, x_4,... on the same t.s. x_t.

Table 5

Time series "Addis Abeba Teff Arrivals (in quintals) 1976-1978"
Centered 4-term moving average and 5-term moving average

x	Time series: x_t	4-term moving average:	Centered 4-term moving averages	5-term moving average:
1	162.6	=	=	=
		=	=	=
2	175.5	=	=	=
		167.5	=	=
3	183.7		169.3	169.3
		171.0		
4	148.1		169.1	168.8
		167.1		
5	176.8		161.5	161.5
		155.9		
6	159.7		156.5	155.3
		157.1		
7	139.0		157.8	162.0
		158.4		
8	152.8		160.9	162.8
		163.5		
9	181.9		159.2	151.7
		154.9		
10	180.4		150.1	146.8
		145.3	=	=
11	104.4	=	=	=
12	114.5	=	=	=

Centered moving averages column 4 were calculated as averages of every two consecutive noncentered moving averages (from column 3). Alternative method of calculation is presented below in Table.

Working Table

Calculation of 4-term centered moving averages (part of time series from Table 5)

Time Series x_t	Sum of 4consecutive t.s.observations	Sum of pairs from column 2	Centered n.average (Column3):8
1	2	3	4
162.6			
175.5			
	669.9		
183.7		1354.0	169.3
	684.1		
148.1		1352.4	169.1
	668.3		
176.8		1291.9	161.5
	623.6		
159.7		1251.9	156.5
	628.3		
139.0			
152.8			

2.5. SEASONAL VARIATION AND SEASONAL ADJUSTMENT

2.5.1. Definition

The seasonal variation is composed of regular, repeated changes around trend line. It arises from the changing seasons and from the rhytm of human activity connected with seasons.

In practice the above definition of seasonal variation is extended to include all repetitive variations not only produced by seasons but also of organizational, administrative and legislative nature.

The strict seasonal variations are very strong in the countries of the temperate climate zone, i.e. in the countries with a strongly differenciated seasons of the year, with frosty and snowy winters and rather warm summers. Strongly influenced by climatic conditions are: house and industrial building, transport and all industries working on the oper area, such as some kinds of mining. There are some industries strictly connected with supplying of the seasonal row materials. An example of extremely seasonal industry is production of sugar which is performed in the course of three months, in the time of so called sugar campaign.

Strong seasonality occurs in time series of external and internal trade. There are some commodities which are purchased and sold in the constant sub-periods of the year (seasonal fruits and vegetables, clothes, etc.)

There are some phenomena for which seasonal fluctuations indirectly follow from the seasons of the year. For example traffic of passangers by almost

all means of transport is connected with seasons of the year. There are seasonal peaks in the tourist traffic etc.

As it was mentioned above fluctuations in statistical data follow from organization of the production and social life and from legislative regulations. For instance different payments which people have to bring (house rents, telephone, taxes) are not-equally distributed in time. Usually majority of the payments are accomplished in the last days of the payment period.

In some countries with planned economy industrial production is distributed unequally. If a factory has to implement an annual plan which is divided on quarter subplans and if this factory has to settle account from each quarter subplan it is self-understanding that all delays from the first and second months are made up in the last month of the quarter.

Of course it does not mean that unrythmical production is a constant attribute of the centrally planned economy. Such so called pursuit of the plan is characteristic for the underdeveloped economics and is caused by the weekness of the infrastructure (transport) and uneffective organization.

It seems that problem of seasonality in Ethiopian time series with short-term observations is more complicated than that for the countries from temperate climate zone. The seasons of year in Ethiopia are not very strongly differenciated. Of course rainy season can cause some troubles with transport and can give rise to demand for some kind of clothes. Supply and demand for different agricultural commodities varies according to periods of year and so on. It seems that in t.s. of developing countries there

is a strong irregular variability. Sometimes random variability is so big that seasonal pattern is completely dominated by it and in consequence seasonality is difficult to reveal and estimate.

2.5.2. Reasons for seasonal variation estimation

In some situations the seasonal variation may be a serious nuisance requiring special statistical attention. For example an economic policy maker who is trying to determine the underlying movement in a t.s. may find that on quarter to quarter basis the longer term trend or cyclical movements are completely dominated by the seasonal variation (unreadable).

In such a case it may be desirable to seasonally adjust the t.s., that is to eliminate the seasonality from t.s., in order to emphasize the longer run components.

Another reason for seasonal variation estimation is analysis of the seasonal pattern for planning and forecasting purpose. Knowledge about the structure of seasonal variations is very important for the proper supplying in seasonal goods, fuel and energy and for proper organization of the transportation service and security. In many situations knowledge or possibility to forecast of seasonal ups and downs enables to avoid big losses.

2.5.3 Kinds of the t.s. components connections

Time series may be concerned as the sum of trend, seasonality and irregular term, i.e.

$$X = T + A_m + I \tag{2.39}$$

or as the product of factors

$$X = T \times S_m \times I \tag{2.40}$$

The former model is the sum of mutually independent components and is called t.s. additive model. From (2.39) it is evident that trend, seasonality and irregular variations are generated independently each other. As the seasonality is determined as repeated up and down variability around the trend it is assumed that yearly sum of the seasonal deviations from trend equals zero.

In the second model trend is corrected by seasonal and irregular factors. This model is called a multiplicative t.s. model.

Seasonal components from this model can be presented as

$$S_m = (1 + b_m) \tag{2.41}$$

b_m can be negative or positive, m means that both S_m and b_m are measures of seasonality connected with subperiods of year, i.e. with months in t.s. with monthly observations, with quarters in t.s. with quarterly observations and so on. Then taking into consideration (2.41), model (2.40) can be presented as follows

$$X = T(1 + b_m) \times I \tag{2.42}$$

and

$$X = (T + T\,b_m) \times I \tag{2.43}$$

where $\sum b_m = 0$ for each year,

Tb_m is seasonal variation presented in the way alternative to (2.40), as a time series of absolute seasonal deviations from trend.

Tb_m are partly determined by trend, partly by seasonality of t.s. Therefore, our conclusion is that seasonality in the model (2.40) is not independent from trend.

In practice models with interdependent trend, seasonality and irregular terms seems to be closer to the real economic and social life than additive models because it

seems to be more justified that size of seasonal and irregular variations varies according to the changes of t.s. observations level.

Now we are going to explain the simplest method of t.s. decomposition. We assume that we are dealing with t.s. with monthly observations.

2.5.4. Time series decomposition by means of step-by-step method (multiplicative model)

To estimate seasonal movements in monthly data we should determine the ratios of these data to the corresponding values of the centered 12 month moving average computed beforehand or to the trend values calculated on the base of analytical function adjusted to the t.s. As result we receive t.s. of ratios

$$R_m = X/\hat{T} = T \times S_m \times I/\hat{T} = S_m \times I \times D \quad (24.4)$$

where D is contamination introduced to the R_m ratios because estimated trend ($\hat{T}$) usually is not quite adequate to the time series data.

When simple mathematical function or moving average is applied for estimation of trend often there are some parts of time series for which estimated trend do not pass exactly thrugh dispersion of data. Simple mathematical function or moving average may prove not sensitive enough to reproduce all fluactuations of the time series. It is not drawback of that trend estimates since according to our definition trend is an image of long term general direction of development. Therefore, fluctuations defined as good and bad years could be interpreted as random deviations from the trend lasting few

month or even total years. Looking from this point of view D stands among others for this long run random deviations from trend.

From (2.44) is evident that for calculation of S_m coefficients we should eliminate IxD factor. To receive monthly coefficients (S_m), ratios R_m should be averaged separately for each subperiod (month) m simply by calculating arithmetic mean for each monthly subset of ratios.

The averages (sometime expressed in percentages) obtained are called Preliminary Seasonal Indexes and represent $S_m D = S'_m$, The average of the seasonal indexes S_m would equal to one (or 100%). It is among others because of the D' component inclusion that average of the preliminary seasonal indexes do not satisfy this condition.

Therefore, to adjust preliminary seasonal indexes to give average 1 (or 100%) we should calculate arithmetic mean of the S'_m

$$\bar{S} = \frac{1}{12} \sum_{m=1}^{12} S'_m = D' \qquad (2.45)$$

and divide each S'_m into $\bar{S}$

$$S'_m / \bar{S} = S_m D' / D' = S_m \qquad (2.46)$$

to receive seasonal indexes S_m whose average equals 1(or 100%)

Additive model

If we consider seasonal movements to be independent of the trend movements i.e. as an additive component rather than as a multiplicative factor, we may construct seasonal indices by computing the differences between original

monthly data and the corresponding values of the 12- month moving average or trend values calculated on the base of function adjusted to the t.s. data, i.e.

$$R_{mj} = X - \hat{T} = (T + A_m + I) - \hat{T} = A_m + I + D \qquad (2.47)$$

where j= 1, 2, ..., N - numbers of consecutive years. To eliminate irregular component and part of the contamination D we should average out of the differences X - T for each month separately over a number of years

$$A'_m = \frac{1}{N} \sum_{j=1}^{N} R_{mj} \text{ for } m = 1, 2, \ldots, 12 \qquad (2.48)$$

Values A'_m are <u>preliminary seasonal components</u>.

Each A'_m is composed of true monthly seasonal component A_m and residual contamination D'.

Seasonal components A_m, as up and down deviations from trend, should fulfil condition

$$\sum_{m=1}^{12} A_m = 0 \qquad (2.49)$$

To estimate A_m measures satisfying above condition we should calculate average value of the preliminary seasonal components A'_m and subtract it from each A'_m value

$$\bar{A} = \frac{1}{12} \sum A'_m \qquad (2.50)$$

$$A_m = A'_m - \bar{A} \qquad (2.51)$$

seasonal deviations A_m (absolute values) can be presented in the form of seasonal coefficients (relative measures) by calculation ratios of the seasonal deviation A_m to corresponding trend value from the examined year and by addition of 1 to each ratio, i.e.

$$S_{mj} = 1 + A_m/T_{mj} \qquad (2.52)$$

From (2.52) follows that for each year we get one set of twelve seasonal coefficients S_{mj}. Each S_{mj} shows what parts of the trend value from year j, month m, represents seasonal deviation A_m

Exercise 3

Estimate trend and multiplicative seasonal factors for the time series with quarterly observations "Volume of exports of pulses in metric tons in 1973 - 1978"
Apply as approximation of a trend

1. A simple mathematical function
2. Moving average

Table 6

Volume of exports of poulses in metric tons in 1973-1978

Years	Quarters				Totals for years
	I	II	III	IV	
1973	34.7	41.3	29.1	36.5	141.6
1974	53.2	24.4	23.7	25.3	126.6
1975	31.5	31.8	18.8	16.7	98.8
1976	17.7	32.9	19.1	21.2	90.9
1977	27.4	24.9	18.3	15.0	85.6
1978	9.1	3.8	3.0	8.0	23.9

Source: Quarterly Bulletin [24] , June 1979

The choice of trend was done for straight line, exponential, power and logarithmic curves. We get the following coefficient of determinations (R^2):

straight line	0.634
exponential	0.603
logarithmic	0.506
power	0.408

The best fitting is a straight line from which the following equation was estimated

$$x_t = 40.37 - 1.34t. \qquad (2.53)$$

If we are forced to perform calculations by hand or only with simple (unprogrammable) calculator we can apply simplified approach which requires for less computational work. Instead of fitting curve to the original time series data, we can adjust trend line to the annual averages of the time series observations using as time variable corresponding averages of original time variable t. We should receive parameters not much different from these resultting from employment of full set of data. We receive following equation of a straight line

$$\hat{x}_t = 39.7 - 1.28t \qquad (2.54)$$

Straight line trends resulting from both (2.53) and (2.54) equations are shown as columns 3 and 4 in working Table 1. Simplified estimation procedure is shown in Working Table 2 where logarithmic curve is adjusted to time series of annual averages.

Working Table 1

t_j	Time series x_t	Linear trend $\hat{x}_t$ (equ.2.53)	Linear trend $\hat{x}_{t1}$ (equ.2.54)	Ratios $x_t/\hat{x}_t$	Centred 4-term m. av $\bar{x}_t$	Ratios $x_t/\bar{x}_t$
1	34.7	39.0	38.4	0.8897		
2	41.3	37.7	37.1	1.0955		
3	29.1	36.4	35.9	0.7995	37.7	0.7715
4	36.5	35.0	34.6	1.0429	37.9	0.9631
5	53.2	33.7	33.3	1.5786	35.2	1.5114
6	24.4	32.3	32.0	0.7554	33.1	0.7372
7	23.7	31.0	30.7	0.7645	29.0	0.8172
8	25.3	29.7	29.5	0.8519	27.2	0.9301
9	31.5	28.3	28.2	1.1131	27.5	1.1455
10	31.8	27.0	26.9	1.1778	25.8	1.2326
11	18.8	25.6	25.6	0.7344	23.0	0.8174
12	16.7	24.3	24.4	0.6872	21.4	0.7804
13	17.7	23.0	23.10	0.7696	21.6	0.8194
14	32.9	21.6	21.8	1.5231	22.2	1.4820
15	19.1	20.3	20.5	0.9409	24.0	0.7958
16	21.2	18.9	19.3	1.1217	24.2	0.8760
17	27.4	17.6	18.0	1.5568	23.1	1.1861
18	24.9	16.3	16.7	1.5276	22.2	1.1216
19	18.3	14.9	15.4	1.2282	19.1	0.9581
20	15.0	13.6	14.1	1.1029	14.2	1.0563
21	9.1	12.2	12.8	0.7459	10.0	0.9100
22	3.8	10.9	11.5	0.3486	6.9	0.5507
23	3.0	9.6	10.2	0.3125		
24	8.0	8.2	8.9	0.9756		
TOTAL	567.4	567.1	568.9	-	-	-

Working Table 2

$\bar{x}_t$ (time series of the averages for years)	$\bar{t}_j$ (Averages of t variable)	$\log \bar{t}_j$	$(\log \bar{t}_j)^2$	$\bar{x}_t \log \bar{t}_j$
34.4	2.5	0.3979	0.1583	14.09
31.6	6.5	0.8129	0.6608	25.69
24.7	10.5	1.0212	1.0428	25.22
22.7	14.5	1.1614	1.3488	26.36
21.4	18.5	1.2672	1.6058	27.12
6.0	22.5	1.3522	1.8284	8.11
141.8	75.00	6.0128	6.6449	126.59

Logarithmic function:

$$X_t = a + \log t + e_t$$

Parameters a and b are calculated on the base of formulas (2.13) and (2.14)

$$b = \frac{126.59 - \frac{141.8 \cdot 6.0128}{6}}{6.6449 - \frac{(6.0128)^2}{6}} = -25.05$$

$$\hat{a} = (141.8:6) - (-25.05)x(6.0128:6) = 48.7$$

Equation of logarithmic curve:

$$X_t = 48.7 - 25.05 \log t \qquad (2.55)$$

For calculation of seasonal coefficients trend calculated on the base of formula (2.53) was applied. This trend is shown in Working Table 1 as column 3 and in column 5 are presented ratios of original observations and corresponding trend values:

$$R_t = X_t/\hat{T}_t = X_t/\hat{x}_t \qquad (2.56)$$

In the Working Table 3 the same ratios are arranged by quarters.

Working Table 3

Years	Quarters			
	I	II	III	IV
1973	0.8897	1.0955	0.7995	1.0429
1974	1.5786	0.7554	0.7645	0.8519
1975	1.1131	1.1778	0.7344	0.6872
1976	0.7696	1.5231	0.9409	1.1217
1977	1.5568	1.5276	1.2282	1.1029
1978	0.7459	0.3486	0.3125	0.9756
Average (S'_m)	1.1090	1.0713	0.7967	0.9637

$$S'_m = 3.9409$$

$$S' = 3.9409:4 = 0.9852$$

Averaging columns of Working Table 3 we received Preliminary Seasonal Coefficients. These preliminary coefficient after adjustment to total 4.000 (average 1.000) constitute Final Seasonal Coefficients S_m. Adjustment is done by dividing each preliminary coefficient S'_m into their average $\bar{S}'$, i.e.

$$S_1 = 1.1090 : 0.9852 = 1.1257$$
$$S_2 = 1.0713 : 0.9852 = 1.0874$$
$$S_3 = 0.7968 : 0.9852 = 0.8088$$
$$S_4 = 0.9637 : 0.9852 = 0.9782$$
$$\text{TOTAL} = 4.0001$$

The same procedure was applied for estimation of constant seasonal coefficients in the case of trend approximated by 4-term centred moving average. Moving average and ratios original to moving average are shown as columns 6 and 7 in Working Table 1. The same ratios arranged by quarters are shown in Working Table 4.

Working Table 4

Quarterly ratios $x_t/\bar{x}_t$.

Years	Quarters			
	I	II	III	IV
1973	-	-	0.7715	0.9631
1974	1.5114	0.7372	0.8172	0.9301
1975	1.1455	1.2326	0.8174	0.7804
1976	0.8194	1.4820	0.7958	0.8760
1977	1.1861	1.1216	0.9581	1.0563
1978	0.9100	0.5507	-	-
Average (S'_m)	1.1145	1.0248	0.8320	0.9212

$$S'_m = 3.8925$$
$$\bar{S} = 0.9731$$

Quarterly Final Seasonal Coefficients are calculated as follows:

S_1 = 1.1145/0.9731 = 1.1453

S_2 = 1.0248/0.9731 = 1.0531

S_3 = 0.8320/0.9731 = 0.8550

S_4 = 0.9212/0.9731 = 0.9467

TOTAL 4.0001

Seasonally adjusted series and goodness of fit

Seasonally adjusted time series is series with eliminated seasonality. Adjustment is performed in order to emphasize the longer run components and to give possibility to follow and interpret irregular deviations from trend.

To eliminate seasonality from time series with multiplicative connections of the trend and seasonality we should simply divide original observations by seasonal coefficients.

Time series with removed seasonality are presented in columns 1 and 4 of Table 5. It is seen that calculated above two sets of seasonal coefficients produced similar seasonally adjusted series.

Seasonally adjusted series X^A contains trend and irregular term:

$$X^A = X/S = TSI/T = TI \qquad (2.57)$$

Since irregulars in multiplicative model are determined as coefficients

$$I_t = (1 + i_t)$$

subtraction of the trend from seasonally adjusted series (2.57) may be shown as estimation of irregular deviations

from trend

$$TI - T = T(1 + i) - T = Ti = \hat{e}_t \qquad (2.58)$$

These irregular deviations for the cases of the straight line trend and moving averages are presented in columns 1, 2 and 5 in Working Table 5 and their squares are shown in columns 3 and 6. Sums of squared irregular deviations ($\sum \hat{e}_t^2$) are applied for calculation of standard errors of the adjusted time series models.

$$S_e = \sqrt{\sum \hat{e}_t^2/(T - k)} \qquad (2.59)$$

where $T - k$ stands for degrees of freedom, k is the number of time series parameters. For the model with straight line trend $k=6$ (2 parameters for straight line trend and 4 parameters for seasonality). For the model with moving average $k=5$ (1 parameter for moving average and 4 parameters for seasonal coefficients).

Working Table 5

Seasonally adjusted time series "Volume of exports of pulses in metric tons in 1973 - 1978" and estimation of its irregular component

The straight line trend			Moving average		
Seasonally adjusted	Residual component	Squarred irregular	Seasonally adjusted	Residual component	Squarred irregular
30.8	-8.2	67.2	30.3	-	-
38.0	0.3	0.1	39.2	-	-
36.0	-0.4	0.2	34.0	-3.7	13.7
37.3	2.3	5.3	38.6	0.7	0.5
47.3	13.6	185.0	46.5	11.3	127.7
22.4	-9.9	98.0	23.2	-9.9	98.0
29.3	-1.7	2.9	27.7	-1.3	1.7
25.9	-3.8	14.4	26.7	-0.5	0.3
28.0	-0.3	0.1	27.5	0.0	0.0
29.2	2.2	4.8	30.2	4.4	19.4
23.2	-2.4	5.8	22.0	-1.0	1.0
17.1	-7.2	51.8	17.6	-3.8	14.4
15.7	-7.3	53.3	15.5	-6.1	37.2
30.3	8.7	75.7	31.2	9.0	81.0
23.6	3.3	10.9	22.3	-1.7	2.9
21.7	2.8	7.8	22.4	-1.8	3.2
24.3	6.7	44.9	23.9	0.8	0.6
22.9	6.6	43.6	23.6	1.4	2.0
22.6	7.7	59.3	21.4	2.3	5.3
15.3	1.7	2.9	15.8	1.6	2.6
8.1	-4.1	16.8	7.9	-2.1	4.4
3.5	-7.4	54.8	3.6	-3.3	10.9
3.7	-5.9	34.8	3.5	-	-
8.2	0.0	0.0	8.5	-	-
564.2	-4.7	840.4	563.1		426.7

Seasonally adjusted series (columns 1 and 4 in Working Table 5)were calculated by dividing original t.s. observations into estimated seasonal coefficients. Therefore, seasonally adjusted series contains trend and irregular term:

$$X^A = X/S_m = TS_mI/S_m = TI \qquad (2.60)$$

Substructing from X^A series estimated trend (moving average) we received estimation of irregular deviations from trend (columns 2 and 5).

$$X^A - \hat{T} = T\,I - \hat{T} = T(1+i_t) - \hat{T} = T + Ti_t - \hat{T} = \hat{e}'_t, \qquad (2.61)$$

which are the base for calculation of standard errors of the adjusted time series models

1) Goodness of fit for the time series model with the straight line trend is as follows:

i) Standard error

$$S_e = \sqrt{840.4/18} = \sqrt{46.7} = 6.8$$

Standard deviation and variance of t.s.

$$S(x) = 10.9 \quad S^2(x) = 118.9 \quad \sum(X_t - \bar{X})^2 = 2854$$

ii) Determination coefficient

$$R^2 = 1 - 840.4/2854 = 0.706$$

About 70 % of the total time series variance is explained by adjusted trend and seasonal coefficients.

2) Goodness of fit for the time series model with moving average trend is as follows:

i) Standard error

$$S_e = \sqrt{426.7/15} = \sqrt{28.4} = 5.3$$

Standard deviation and variance of t.s. (without first and last two observations)

$$S(x) = 10.6 \qquad S^2(x) = 112.7 \qquad \sum(x_t - \bar{x})^2 = 2254$$

ii) Determination coefficient

$$R^2 = 1 - 426/2254 = 0.811$$

About 81 % of the total time series variance is explained by adjusted moving average and seasonal coefficients.

2.5.5. Changes (shifts) in seasonal pattern

So far we were dealing with constant seasonality i.e. seasonal variation was estimated as one set of monthly or quarterly seasonal seasonal coefficients or deviations from trend, which served as a measure of seasonal variation for all examined period. But in the real life seasonality undergo to changes, such as changes of seasonal amplitudes i.e. size of deviations from trend or/and shifts of seasonal peaks from one to another month. Changes of seasonal amplitudes are shown in Fig.13 shift of seasonal peak is shown in Fig.14. These two kinds of changes may occur jointly.

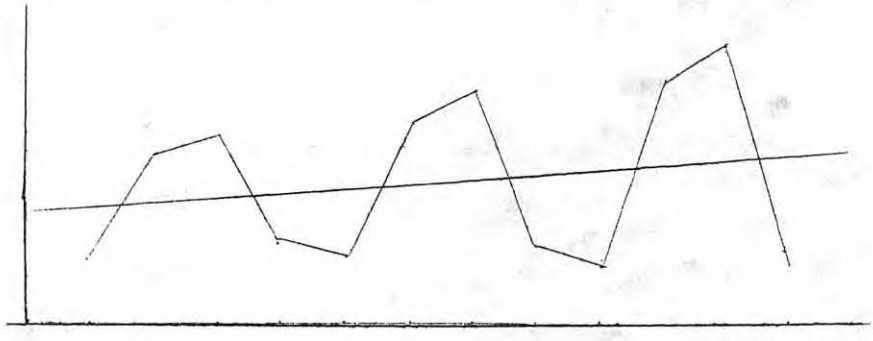


Fig.13. Seasonality with increasing amplitudes

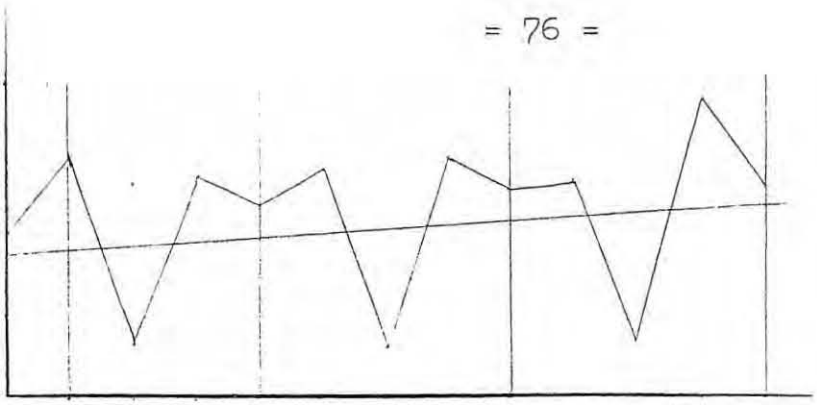


Fig.14. Seasonality with shifted seasonal peaks.

Changes of the structure of seasonality may be abrupt, irregular or progressive (regular).

The reason for abrupt change may be war, change of organization or legislative regulations. This kind of change consists in the shift or even in the reversing of the seasonal ups and downs.

The best what can be done when we are dealing with such t.s. is to analyse each part (before and after change) separately.

Seasonality with irregular changes of seasonal pattern, i.e. random shifts of seasonal peaks and/or random changes of amplitudes of seasonal deviations from trend, should be regarded as seasonality with constant pattern. However, in the case when irregular disturbances are very strong is difficult to separate seasonality from irregular term.

So called evalutionary changing seasonal structure, i.e. progressively increasing (decreasing) seasonal amplitudes and/or gradually shifting seasonal peaks is an example of regular changes. This kind of change is characteristic for developing economies. The reason for such changes are in-

vestments, technical innovations, improvement in management and organization of economy, production and social life.

The simplest case of the changing seasonality are seasonal deviations from trend, increasing (decreasing) proportionally to the increase (decrease) of a trend. Such deviations can be presented as constant seasonal coefficients in multiplicative t.s. model. It means that if we want to present multiplicative t.s. with constant seasonal coefficient by means of additive model we will receive additive relations with seasonal deviations from trend increasing (decreasing) proportionally to the increase (decrease) of the trend.

On the other hand if the true constant additive deviations from trend are presented by means of multiplicative model, i.e. as coefficients correcting trend, we receive time series of changing seasonal coefficients.
Very often seasonality changes independently from trend or it only partly depends on the trend. It means that after estimation of trend, we should estimate monthly (quarterly) seasonal coefficients or deviations as simple functions of time. For each month (quarter) we have got separate function which enables to calculate seasonal coefficient (deviation) for each year. We can extrapolate measure of seasonality beyond the range of time series, for instance one year ahead.

2.5.6. Step-by-step method for changing seasonality

Application of step-by-step approach for estimation changing seasonality is shown in exercise below.

Exercise 4

Estimate seasonal variation for "Sugar production in Ethiopia in 1972-1978".

Apply additive model. Choose between constant and changing seasonality.

Table 7

Sugar Production in Ethiopia (in 100 kg)

Years	Quarters				Total
	I	II	III	IV	
1972	443.4	357.6	4.0	401.8	1,206.8
1973	531.1	377.4	2.1	330.4	1,241.0
1974	351.7	438.5	31.4	386.7	1,208.3
1975	540.9	345.4	14.3	349.0	1,249.6
1976	542.0	327.0	14.9	363.9	1,247.8
1977	463.2	396.4	17.6	369.4	1,246.6
1978.	511.8	536.9	30.5	493.4	1,572.6

Source: Quarterly Bulletin, Economic Research and Planning Division, Addis Ababa.

Working Table 1

Adjustment of a straight line trend to the annual averages

Years	Averages $\bar{x}_j$	t_j	$t_j - \bar{t}$	calculated trend $\hat{x}_j$
1972	301.7	2.5	-12	289.7
1973	310.3	6.5	- 8	300.0
1974	302.1	10.5	- 4	310.2
1975	312.3	14.5	0	320.5
1976	312.0	18.5	4	330.7
1977	311.7	22.5	8	341.0
1978	393.2	26.5	12	351.2

The straight line adjusted to data from column 2 of Working Table 1, using time variable t_j (column 3) is as follows

$$X_j = 283.3+2.56\ t_j \qquad (2.63)$$

Determination coefficient $R^2 = 0.468$

To calculate trend values corresponding to original time series observations one should apply equation

$$X_t = 283.3 + 2.56t \qquad (2.64)$$

where t is time variable from Working Table 2.

Working Table 2

Trend adjusted to the quarterly data, and differences $x-\hat{x}$

t	Original t.s x_t	Trend $\hat{x}_t$	Differences $x-\hat{x}$
1	443.4	285.9	157.5
2	357.6	288.4	69.2
3	4.0	291.0	-287.0
4	401.8	293.5	108.3
5	531.1	296.1	235.0
6	377.4	299.0	78.4
7	2.1	301.2	-299.1
8	330.4	303.8	26.6
9	351.7	306.4	45.3
10	438.5	308.9	129.6
11	31.4	311.5	-280.1
12	386.7	314.1	72.6
13	540.9	316.6	224.3
14	345.4	319.2	26.2
15	14.3	321.8	-307.5
16	349.0	324.3	24.7
17	542.0	326.9	215.1
18	327.0	329.4	- 2.4
19	14.9	332.0	-317.1
20	363.9	334.6	29.3
21	463.2	337.1	126.1
22	396.4	339.7	56.7
23	17.6	342.3	-324.7
24	369.4	344.8	24.6
25	511.8	347.4	164.4
26	536.9	349.9	187.0
27	30.5	352.5	-322.0
28	493.4	355.1	138.3
TOTAL	8972.7	8973.4	- 1.0

Working Table 3
Differences arranged by quarters

Years	Additive model differences			
	I	II	III	IV
1972	157.5	69.2	-287.0	108.3
1973	235.0	78.4	-299.1	26.6
1974	45.3	129.6	-280.1	72.6
1975	224.3	26.2	-307.5	24.7
1976	215.1	- 2.4	-317.1	29.3
1977	126.1	56.7	-324.7	24.6
1978	164.4	187.0	-322.0	138.3
Arith. average	166.8	77.8	-305.4	60.6

Since total of preliminary additive deviations from trend is

$$A'_m = -0.2 \simeq 0$$

we can take values A'_m as estimates of final constant seasonal deviations, i.e.

$$A_1 = 166.8; \quad A_2 = 77.8; \quad A_3 = -305.4; \quad A_4 = 60.6$$

We have found linear functions for each of the four subsets of differences from Working Table 3. These functions are the base for calculation of the preliminary seasonal coefficients and preliminary seasonal deviations. They are shown below:

For the first quarter

$$A'_{1t} = 170.7 - 1.0t \qquad R^2 = 0.01$$

For the second quarter

$$A'_{2t} = 52.4 + 6.4t \qquad R^2 = 0.05$$

For the third quarter

$A'_{3t} = -277.8 - 6.9t$ $\quad R^2 = 0.74$

For the fourth quarter

$A'_{4t} = 54.5 + 1.5t$ $\quad R^2 = 0.005$

where t = 1, 2, 3, 4, 5, 6, 7.

For projecting (forecasting) seasonal deviations one year ahead t = 8 should be substituted to the adjusted functions.

The preliminary changing seasonal deviations calculated on the base adjusted functions are shown in Working Table 4.

Working Table 4

Preliminary seasonal components A'_{mt}

Year	Quarters				Totals
	I	II	III	IV	
1972	169.7	58.8	-284.7	56.0	-0.2
1973	168.7	65.2	-291.6	57.5	-0.2
1974	167.7	71.6	-298.5	59.0	-0.2
1975	166.7	78.0	-305.4	60.5	-0.2
1976	165.7	84.4	-312.3	62.0	-0.2
1977	164.7	90.8	-319.2	63.5	-0.2
1978	163.7	97.2	-326.1	65.0	-0.2

Since total of preliminary seasonal deviations is nearly zero, for each year, we can take them as the good approximation of the final seasonal deviations, i.e.

$$A'_{mt} = A_{mt}$$

Seasonally adjusted original series and estimated irregular term are shown in Working Table 5.

Working Table 5

Additive changing seasonality			Additive constant-seasonality		
Seasonally adjusted t.s.	Irregular term	Squarred irregulars	Seasonally adjusted t.s	Irregular term	Squared irregular
273.7	-12.2	148.8	276.6	-9.3	86.4
298.8	10.4	108.2	279.8	-8.6	74.0
288.7	- 2.3	5.3	309.4	18.4	338.5
345.8	52.3	2735.3	341.2	47.7	2275.3
362.4	66.3	4395.7	364.3	68.2	4651.2
312.2	13.2	174.2	299.6	0.6	0.4
293.7	- 7.5	56.3	307.5	6.3	39.7
272.9	-30.9	954.8	269.8	-34.0	1156.0
184.0	122.4	14981.8	184.9	-121.5	14762.3
366.9	58.0	3364.0	360.7	51.8	2683.2
329.9	18.4	338.6	336.8	25.3	640.0
327.7	13.6	185.0	326.1	12.0	144.0
374.2	57.6	3317.8	374.1	57.5	3306.3
267.4	-51.8	2683.2	267.6	-51.6	2662.6
319.7	- 2.1	4.4	319.7	- 2.1	4.4
288.5	-35.8	1281.6	288.4	-35.9	1288.8
376.3	49.4	2440.4	375.2	48.3	2332.9
242.6	-86.8	7534.2	249.2	-80.2	6432.0
327.2	- 4.8	23.0	320.3	-11.7	136.8
301.9	-32.7	1069.3	303.3	-28.7	823.7
298.5	-38.6	1490.0	296.4	-40.7	1656.5
305.6	-34.1	1162.8	318.6	-21.1	445.2
336.8	- 5.5	30.3	323.0	-19.3	372.5
305.9	-38.9	1513.2	308.8	-36.0	1296.0
348.1	0.7	0.5	345.0	- 2.4	5.8
439.7	89.8	8064.0	459.1	109.2	11924.6
356.6	4.1	16.8	335.9	-16.6	275.5
428.4	73.3	5372.9	432.8	77.7	6037.3
8974.1	0.7	63451.6	8974.1	3.3	65851.4

Summary results:

i) Totals of the seasonally adjusted series for both changing and constant seasonality are 8974.1.

ii) Variance of the original series: 36334 $\sum (x_t - \bar{x})^2 = 1,017,352.$

iii) Total of the squared irregular term

- for changing seasonality: 63,451.6
- for constant seasonality: 65,851.4

iv) Number of degrees of freedom

- for model with changing seasonality: 28-10=18 (2 trend parameters 4x2 seasonal parameters),
- for model with constant seasonality: 28-6=22 (2 trend parameters +4 seasonal parameters).

v) Standard errors

- for changing seasonality model

$$S_e = \sqrt{63,451.6/18} = \sqrt{3525.1} = 59.4$$

- for constant seasonality

$$S_2 = \sqrt{65,851.4/22} = \sqrt{2993.2} = 54.7$$

vi) Determination coefficients

- for changing additive seasonality

$$R^2 = 0.938$$

- for constant seasonality

$$R^2 = 0.935$$

Conclusion

Model with the straight line trend and constant additive seasonality explains about 93 % of the t.s. variance while model with the same trend and changing additive seasonality explains about 94 % of the t.s. variance.

2.5.7. Separation of the business cycle component

Time series decomposition methods were originated around the end of XIX th century and were results of works attempting to separate from time series business cycle indicators.

Big advance in decomposition methods was done at the beginning of the XXth century when was developed and

generalized the procedure of trend estimation by means of simple and weighted moving averages. Same economist tried to isolate and predict changes of business cycle. To do it they had to estimate and remove trend and seasonal variation. In France a committee was appointed which in 1911 presented a report dealing with the reasons and causes of economic crisis of 1907. It introduced the idea of so called leading and coincidental business cycle indicators and attempted to separate the trend from the cycle so that the movement of the latter could be followed. At the same time the method was elaborated upon in the USA, and the idea of constructing so called barometers of business activity was developed. Barometers of business activity cinsisted of the set of leading and coincidental indicators and method of trend and seasonal factors separation from the time series.

The process of decomposition was refined by F.R. Macaulay [13] who in 1930 introduced the ratio-to-moving-averages method, variations of which are the most widely used today. It consists of three basic steps: first, the calculation of the preliminary seasonal factors, which are the ratios of the original observations to a 12 month moving averages, then through averaging the seasonal factors we obtain 12 monthly (or 4 quarterly) seasonal indices; secondly, the trend is calculated for each point and thirdly, the trend is divided into the moving averaged data with the result of obtaining the cyclical factors. That is, assumming a multiplicative relationship (an additive model will behave in a similar way), whose seasonal pattern is of 12- period duration, we have:

$$X = T \cdot C \cdot S \cdot I \qquad (2.65)$$

where T is the trend component,
C is the cyclical component
S is the seasonal component
I is the random (irregular) component

The first step consists of

$$M = TC \tag{2.66}$$

where M is a 12- term moving average of X which eliminates the seasonality and a great deal of irregular variability. Next, dividing original t.s. into moving average M we get preliminary seasonal coefficients

$$S' = X/M = TCSI/TC = SI \tag{2.67}$$

Through averaging (2.67) we eliminate the irregularity, so the remaining can be considered as the seasonal indices:

$$S_j = \frac{1}{12} \sum_{i=1}^{12} S'_{ij} \tag{2.68}$$

$j = 1, 2, \ldots, N$

N is the number of years.

Step two is

$$T = \varphi(t) \tag{2.69}$$

where φ is function (a straight line, exponential, logarithmic and so on) fitted to the original data by the least squares method.

Finally, in step three we divide (2.69) by (2.66) to obtain C, an estimate of the cyclical element

$$C = TC/T \tag{2.70}$$

The cyclical fluctuations are long term fluctuations with random period and amplitude, i.e. the length of the cycles and distance from the peak to the bottom vary randomly.

For the analysis of the present state of the cycle, i.e. for the analysis if at present time cycle is in its way down, up or is changing its direction, we should apply estimated seasonal factors and predicted trend value.

Dividing actual t.s. observations by corresponding seasonal coefficients and forecasted trend we get data about actual state of cyclical fluctuation i.e.

$$X/TS = CI \qquad (2.71)$$

Information about the cyclical movement is contaminated by irregular variation. Usually it is possible to assess the direction of the cyclical changes even in the presence of the irregular disturbances.

Sometimes, seasonally adjusted data, i.e. data containing the trend, cycle and irregular variation is presented instead of cycle - irregular data. Such data show both trend and cyclical changes.

2.5.8. Computer programs for the t.s. decomposition

Presented above the ratio-to-moving-averages method is the basis for computerized method of the t.s. decomposition known as the Census II Method, which has been developed in USA and now is used all over the world.

This method (see publications [4, 14, 20] differs from Macaulay's original method in that (a) it uses improved moving averages in order to compute the seasonal factors and the trend-cycle curves (b) it replaces the values lost in the beginning and the end of the series because of the moving averages by calculated values, (c) it replaces extreme (outlying) observations by some assessed values, (d) it takes into account trading day variation ,

(e) most important, it isolates each component one at a time, by first calculating crude estimates which are further refined.

This refinment process consists of (a) the estimation of the trend-cycle by means of 12-months m.av. and by the calculation seasonal factors on the base of ratios to m.average, (b) elimination of the seasonal variation from original t.s., (c) the estimation of improved trend-cycle component by means of weighted m.av., such as 13-term Spencor's m.av., (d) calculation of the final seasonal factors, starting from the ratios to weighted moving average (point c above).

Simplified version of the Census II method was programmed in the Computer Centre of the Central Statistical Office. This program is adapted for the t.s. with monthly observations, no longer than 12 years.

The output of the CSO computer program:

i) Preliminary seasonal factor
ii) Adjusted seasonal factors (to give total of 12 for each year).
iii) Rate of growth of the 12-months moving average (trend-cycle).
iv) Seasonally adjusted series.
v) Seasonally adjusted series smoothed by 3-months m.av.

2.6. APPLICATION OF THE TIME SERIES DECOMPOSITION RESULTS TO STATISTICAL ANALYSIS

2.6.1. Analysis of seasonally adjusted time series

Seasonal fluctuations very often make impossible to assess present direction of the changes in analysed phenomena. To make successive time series observations comparable one has to remove seasonal variation from them.

For the elimination of seasonality from actual data one should use seasonal coefficients or seasonal deviations extrapolated for the actual year.

If changing seasonal variation is approximated by set of functions

$$S_{ij} = \varphi i(j_i\ e_j) \qquad (2.72)$$

where i denots month (or quarters), i.e.

i= 1, 2, ..., 12 (or i= 1, 2, ..., 4)

j= stands for consecutive years, j= 1,2,..., M, then extrapolated (projected) seasonality is calculated as

$$S_{iM+1} = \varphi i\ (M+1) \qquad (2.73)$$

If seasonality is assumed to be constant then set of constant seasonal coefficients (deviations) should be applied for the present year.

Removing of seasonality is done simply by dividing original data into seasonal coefficients or by subtracting seasonal deviations.

There are many time series for which observation of the actual trends is very important. For example trends of retail and wholesale prices of agricultural goods may inform about changes of their supply (reserves). Fast increase of prices may indicate shortage of commodity in a country or in a region. Changing prices of seeds during sowing time may indicate changes of farmers preferences for crops. On the other hand actual market prices of crops are important factor influencing farmers decision about area allocated for these crops. From the point of view of economic policy-maker trends of stock reserves for food and trends of actual prices for food commodities on international markets are very important.

The seasonally adjusted data gives only first rough impression of possible change and should be stimulus for further statistical and economic analysis which would indicate if observed changes are permanent tendency or only temporary random fluctuation.

Another advantage of actual seasonally adjusted data analysis is that it can show how effective are measures taken by economic manager to avoid unfavourable situation.

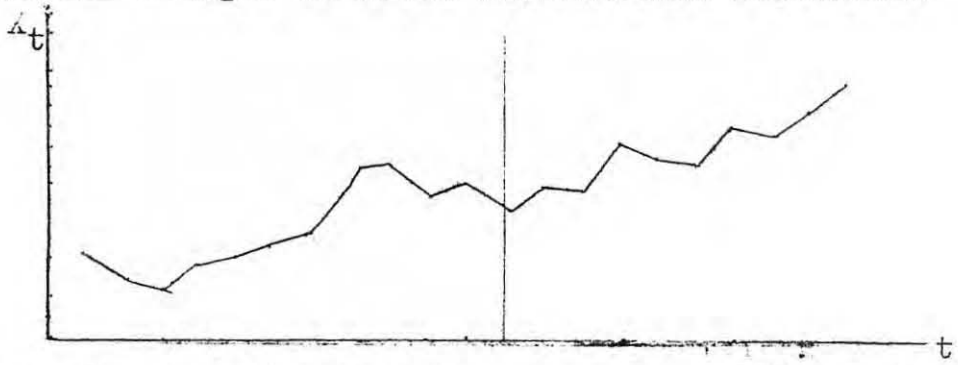


Fig.15. Food Retail Price Index for Addis Ababa. Original data.

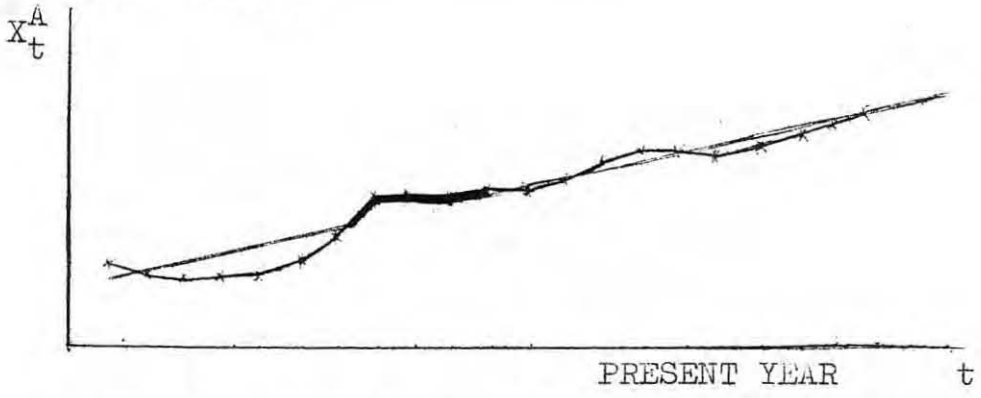


Fig.16. Food Retail Price Index for Addis Ababa.
~~x-x-x-~~ seasonally adjusted data
_____ trend.

Fig.16. shows original data for Food Retail Price Index for Addis Ababa, in the previous and present years. We, cannot assess if a big increase of the price index in months August-September of the present year should be regarded as result of seasonality or as result of general increasing trend.

Fig.17. shows the trend and data after removing seasonal variation. This diagram confirms that we are dealing with real increase of the price index.

The comparison of estimated trend with original seasonally adjusted data enables one to assess whether the development of the examined process is consistent with the forecasted trend.

At present,possibilities for making seasonal adjustment on the data for prices of food are rather limited. Data on prices of major grains were collected by Food and Nutrition Surcillance Programme/Relief and Rehabilitation Commision (compare [23]) since 1977. These data are collected monthly from almost all the agriculturally important weredas of the country. However these time series are too short to estimat seasonality and to perform seasonal adjustment. It should be borne in mind that for many regions and crops prices show big irregular fluctuations due to big fluctuations in production, insufficient transport facilities for transporting grains to deficit areas etc.

If random fluctuations are bigger than seasonal this makes the estimation of the latter impossible. However even for time series with irregular variability predominated by seasonality the minimal period for estimation of regular variations is four or five years. So far such long time series

are available for prices of pulses, oilseeds and meat. Price Index for Food for Addis Ababa, Addis Ababa Wholesale Price Indices for Exported Food Commodities and for Imported Food Commodities are also available (see [24]). It seems that development of agricultural price statistics system is necessary for improvement of economic planning and management. This development should be directed to extension of data collection by commodities and regions) and its final stage should be establishment of computerized system of price statistics consisting of data base and methods of statistical analysis including analysis of seasonality, trends and seasonally adjusted series.

2.6.2. Application in the short-term planning

Suppose we have estimated trend and monthly seasonal coefficients of fertilizers supply for agriculture. Having fixed the total amount of the fertilizers supply foreseen for the next year we have to prepare detailed plan of the fertilizers supply in each month of the forthcoming year. In other words we should assess how total annual supply of fertilizers should be destributed over the months of the year to meet demand from state farms, cooperatives and farmers.

Our task could be easily solved if we have got data about the demand in each month of the next year from all users of fertilizers. But if such data are not available or not reliable we should apply statistical method.

The distribution of demand can be estimated on the basis of the forecasted (extrapolated one year ahead) trend and seasonality. Trend determines the development of the process within examined year. Seasonal coefficients (or

deviations) determine the distribution over the months of year.

If we apply multiplicative t.s. model, the product of extrapolated trend and corresponding seasonal factor, i.e.

$$x_{N+i} = T_{N+i} \cdot S_{N+i} \qquad (2.74)$$

where N is the number of the last time series observation, i is the month of the year i= 1,2,...,12,

determines the value for the month i of the year for each plan is prepared.

Having calculated forecasted values for all months of the year, percentage distribution over the months of the examined year could be calculated as follows

$$d_{N+i} = (x_{N+i} / \sum_{i=1}^{12} x_{N+i}) \cdot 100 \qquad (2.75)$$

Example 3 (fictitious data)

1. Suppose we have the following trend function fitted to time series with monthly observations covering period from 1974 to 1978

$$T_t = 115.3 + 0.75t \qquad (2.76)$$

where t = 1,2,..., 60.

2. Extrapolated one year ahead (for 1979) trend is calculated on the ground of equation (2.76) with t= 61, 62, ..., 72 (for example T_{61} = 161.05).

3. Extrapolated monthly seasonal coefficients are calculated on the base of functions adjusted to the subsets of ratios original/trend, arranged by months. If for example function

$$S'_{1t} = 0.9365 + 0.0051t \qquad (2.77)$$
$$t = 1,2,3,4,5$$

was fitted to ratios from Januaries, coefficient for January of the next year should be calculated on the bases of function (2.77) with t=6, i.e. r_{t1}= 0.9365+0.0306 = 0.9671.

In this manner seasonal coefficients for the remaining months should be calculated. If instead of changing (moving) seasonality we received a set of constant seasonal coefficients they should be applied as a set of extrapolated seasonal coefficients.

Suppose that on the base of a set of functions adjusted to the monthly subsets of data (one function for month), we received following extrapolated values of the preliminary seasonal coefficients:

S'_{16} = 0.9671 (January 1979)
S'_{26} = 0.9765 (February 1979)
S'_{36} = 0.9985 (March)
S'_{46} = 1.0221 (April)
S'_{56} = 1.0332 (May)
S'_{66} = 1.0185 (June)
S'_{76} = 1.0385 (July)
S'_{86} = 1.0452 (August)
S'_{96} = 1.0060 (October)

$$S'_{116} = 0.9965$$
$$S'_{126} = 0.9954$$

$$\sum_{i=1}^{12} S'_{i6} = 12.1287, \quad \bar{S}' = 1.0107$$

To adjust coefficients S'_{16} to the total 12, each of them was divided into their average $\bar{S}' = 1.0107$.

Values of trend and adjusted seasonal coefficients are shown in Table 8. Products of trend and seasonal coefficients are shown in the column 4 of the Table. Share of product in the total of products indicates which part of the total forecasted value for 1979 year determines values from this particular month. This shares are presented in the column 5.

Table 8

Values of trend, seasonal coefficients, percentage distribution of the total value for 1979 over the months (d_{N+i}) and cumulative distribution (D_I).

Months/ value of t	Trend for 1979	Seasonal coefficients for 1979	Product (trend · seasonal)	Distribution d_{N+i} (in %)	Cumulative distribution
J/61	161.05	0.9569	154.11	7.77	7.77
F/62	161.80	0.9662	156.33	7.89	15.66
M/63	162.55	0.9879	160.58	8.10	23.76
A/64	163.30	1.0113	165.15	8.33	32.09
M/65	164.05	1.0223	167.71	8.46	40.55
J/66	164.80	1.0077	166.07	8.38	48.93
J/67	165.55	1.0275	170.10	8.58	57.51
A/68	166.30	1.0341	171.97	8.67	66.18
S/69	167.05	1.0203	170.44	8.60	74.78
O/70	167.80	0.9953	167.01	8.42	83.30
N/71	168.55	0.9860	166.19	8.38	91.58
D/72	169.30	0.9849	166.74	8.42	100.00
TOTAL	1982.1	12.0004	1982.4	100.00	-

Cumulative distribution (column 6)

$$D_I = \sum_{i=1}^{I} d_{N+i}$$

shows how many percent of the total yearly plan should be done to the end of the month I. For example D_9 = 74.78 shows that to the end of September should be implemented 74.78% of the total plan for 1979 year.

2.7. BIBLIOGRAPHICAL NOTE

Due to the fast development of time series analysis in the last fifty years there is a very extensive literature on the decomposition approaches and techniques.

The purpose of this biblographical note is to provide supplementary literature on the theory and more sophisticated methods of time series decomposition. Readers wishing to go through systematic, detailed and extensive course of general problems of analysis and decomposition methods should study classical manuals of statistics such as Croxton and Cowden's [5] or Yule and Kendall's [22] . Kendall's classical book [12] is addressed to more advanced readers.

There is extensive literature on particular methods of decomposition for instance Burman [2], Durbin [6] , Durbin and Murphy [7] , Harrison [10] , Rosenblat [17] .

Widely applied computerized Census Method II is presented among others by Shiskin et al. [20], Cleveland and Tiao [4] , Makridakis [14] and by OECD [18].

Review of contemporary time series analysis problems contain books of Chatfield [3] , Brillinger [1] and paper of Makridahis [14] . Methods of business cycle analysis are presented among others by Shiskin [19] , Moore [15] , Moore and Shiskin [16] .

According to autor's knowledge there is no important literature on the application of time series analysis to agricultural surveys.

REFERENCES

1. Brillinger, D.R. (1975). Time Series Data Analysis and Theory. Holt, Reinehart & Winston, New York.

2. Burman, J.P. (1965). Moving seasonal adjustments of economic time series. Journal of the Royal Statistical Society, A, 128, pp534-558

3. Chatfield, C. (1975). The Analysis of Time Series: Theory and Practice. Chapman and Hall, London.

4. Cleveland, W.P. and Tiao, G.C.(1976). Decomposition of seasonal time series: a model for the Census X-11 Program. Journal of the American Statistical Association, 71,581-587.

5. Croxton, F.E. and Cowden, D.J. (1959). Applied general Statistics. Englewood Cliffs.

6. Durbin, J. (1957). Trend elimination for the purpose of estimating seasonal and periodic components. In Rosenblatt, M. (ed), Time Series Analysis. Wiley, New York.

7. Durbin, J. and Murphy, M.J. (1975). Seasonal adjustment based on a mixed additive multiplicative model. Journal of the Royal Statistical Society, Series A, 138, 385-410.

8. Fuller, W.A.(1976). Introduction to Statistical Time Series, John Wiley & Sons, New York.

9. Hannan, E.J. (1963). The estimation of seasonal variation in economic time series. Journal of the American Statistical Association, 58, 31-44.

10. Harrison, P.J. (1965). Short-term forecasting. Applied Statistics, 14, 102-139.

11. Kendall, M.G. (1973). Time Series. Hafner Press, New York

12. Kendall, M.G. The Advanced Theory of Statistics. Vol.II. C. Griffin and Company, London.

13. Macauley, F.R. (1930). The smoothing of time series. National Bureau of Economic Research, pp. 121-136.

14. Makridakis, S. (1976). A survey of time series. International Statistical Review, 44, 29-70.

15. Moore, G.H. (1961). Business cycle indicators. National Bureau of Economic Research.

16. Moore, G.H. and Shiskin, J. (1967). Indicators of business expansions and contractions. National Bureau of Economic Research.

17. Rosenblatt, H.M. (1963). Spectral analysis and parametric methods of seasonal adjustment of economic time series. Proceedings of the Business and Economic Statistics Section, American Statistical Association, pp. 94-133.

or

Bureau of the Census (1970). Technical Paper no.23.

18. Seasonal Adjustment on Electronic Computers. OECD, Paris (1961)

19. Shiskin, J. (1957) Electronic computers and business indicators. National Bureau of Economic Research. Ocassional Paper, no. 57.

20. Shiskin, J. et al. (1967). The X-11 variant of the Census II method seasonal adjustment program. Bureau of the Census, Technical Paper, no.15.

21. Vieira, S. and Hoffman, R. (1977), Comparison of the logistic and Gompertz growth functions considering additive and multiplicative error terms. Applied Statistics, vol. 26, no.2.

22. Yule, G.U. and Kendal M.G. An introduction to the Theory of Statistics. Charles Griffin, London.

23. A Summary Review of the Average Retail Price of Major Grains (January 1977 to May 1979). Food and Nutrition Surveillance Programme Relief and Rehabilitation Commision. Addis Ababa, November 1979 (mimeographed).

24. Quarterly Bulletin (New Series). National Bank of Ethiopia, Research and Planning Division. Addis Ababa.

25. HP-19C/HP-29C Applications Book. Hewlett-Packard Company 1971.

26. HP-67/HP-97 Statistical Package. Hewlett-Packard Company 1976.

27. HP-67/HP-97 Standard Package. Hewlett-Packard Company 1976.

28. Multiple linear regression. (Description of program). Addis Ababa: CSO. (mimeographed).

Chapter III

FORECASTING METHODS

3.1. INTRODUCTION

The main topics of Chapter III is statistical forecasting. We distinguish two classes of forecasting techniques.

A class of forecasting procedures where forecasts of a given variable are based only on the current and past values of this variable will be called univariate or projection methods (compare [4]).
Another class of forecasting is based on attempt to explain the behaviour of forecasted variable by means of econometric model, i.e. through variation in a number of independent (explanatory) variables. The basical concepts of this approach are presented in section 3.5 of the present chapter.

In the last twenty years fast development of projection methods took place. Recent advances in methods of determination of current time series value in terms of its past observations and in terms of its past random (residual) deviations resulted in that some univariate forecasting methods are competitive with predicting on the base of big econometric models, being simultancously easier to utilize and update and in general less complex and cheaper to develop.

Of course this does not imply that econometric models could be substituted by univariate ones. As it will be shown in Section 3.5 econometric models are irreplaceable in respect to simulation of the future situation, i.e.

when several alternatives must be checked and their effect estimated and used for planning purposes.

There are three kinds of prediction techniques within projection approach. The first kind, the eldest, is based on time series decomposition and consists in projection into future estimated trend and seasonality. The second kind of prediction techniques called exponential smothing models consider predicted time series value as weighted average of its current and past observations. Weights of this average usually discard past values at exponential rate. The third kind of prediction, the most complex one is the Autoregressive-Moving Average Scheme, which assumes that a given value of a time series can be adequately described in terms of previous values of the series itself and/or previous error terms.

The development and wide application of highly sophisticated forecasting methods was facilitated by electronic computer techniques.

Time series methods of forecasting may prove very useful for prediction in agriculture particularly for one year ahead forecasts of area under crops and agricultural production. There is a big demand for methods of forecasting able to deal with highly irregular variables and not calling for big amount of supplementary statistical data. It seems that some of the univariate prediction models could be successfully applied to this kind of variables.

We would like to show that methods presented in this chapter may prove useful for prediction at the country level as well as at regional level. Then special attention was paid for adaptation of time series forecasting methods to get along with

time series-cross section data showing simultaneously changes of investigated phenomena in time and its differentiation by regions. It should be stressed that forecasting is not merely performance of algorithms. The significant part of forecasting procedure is conceptual work consisting in the choice of the method and/or model and selection of a period of past data the best for the forecast purpose.

One possible criterion for the selection of the model and period of historical data is goodness of fit measured by means of standard error S_e and coefficient of determination R^2 (compare Chapter 1, Section 1.6 and Chapter 2, Section 2.4). Because of its computational simplicity this criterion is generally applied for assessment the accuracy performance of a forecasting model and in selecting among alternative forecasting methodologies and models. Goodness of fit criterion is recently often criticized (see [14] , page 259) because mathematical models very adequate for historical data sometime give unsatisfactory forecasts.

It should be noted that for forecasting purpose the model which perfectly fit historical data may prove useless as well as a bad fitting one.

It seems that in the case when it is allowed by statistical data a criterion based on a set of predictions made for the periods covered by historical data could be recommended as more justified than criterion based on goodnes of fit. This approach is characterized by Makridahis ([14] , page 259) as follows: "The approach that is gaining wide support at present is to look at a time-series model in terms of its ability to forecast well in a post-sample fashion. That is the 'goodness' of the model must be judged not in terms of how

well it fits historical data, but how well it forecasts, as new values become available which have not been used in developing the forecasting model. There is certainly a difference between post-sample accuracies and those obtained during the model-fitting phase. However one should be aware that this approach can not be used if time series covers short period of past. It seems that for the choice of the length of time series post-sample accuracy could prove useful as additional auxilary criterion.

In this Chapter the following forecasting methods are presented:

1. Simple autoregressive models with one year delayed observations as explaining variables.
2. Adaptation of trend and seasonality projection to forecasting of cross-section-time series data.
3. Holt-Winters Exponential Smoothing and its adaptation to cross-section-time series data.

4. Multiple regression model for agricultural production.

Draft of econometric forecasting model of crop production is presented in section 3.6 . Its purpose is to show the kind of difficulties which arise with use of such models for forecasting and advantages of their successful implementation.

3.2. LINEAR AUTOCORRELATION AND AUTOREGRESSIVE MODELS

Most time ordered observations are not independent of one another. For example level the consumption in month t depends not only on the disposable income but on the level of consumption in months t-1, t-2 etc.

Similarly the amount of produced goods in month t is partly determined by production in months t-1, t-2 etc. Area under crop is partly determined by area under crop in the previous year.

Of course there are random series with mutually independent observations as well as time series without independent observations.

One of the methods used checking whether a given time series is random or not is to examine its autocorrelation coefficients, i.e. correlation between neighbouring observations of a series

$$r_d = \frac{\sum_{t=d+1}^{N} x_t \cdot x_{t-d}}{\sqrt{\sum_{t=d+1}^{N} x_t^2 \cdot \sum_{t=d+1}^{N} x_{t-d}^2}} \qquad (3.1)$$

where $x_t = X_t - \bar{X}_1$

X_t stands for original time series

$x_{t-d} = X_{t-d} - \bar{X}_2$

$\bar{X}_1$ and $\bar{X}_2$ are arithmetic means for X_t and X_{t-d} series respectively.

Formula (3.1) shows that autocorrelation of order d is the correlation between time series X_t and time series composed of observations of the same time series X_t delayed d units of time. For example if d = 1, we calculate corelation coefficient taking time series X_t and the same time series beginning with observation X_2, i.e.

X_t	X_{t-1}
X_2	X_1
X_3	X_2
.	.
.	.
.	.
X_N	X_{N-1}

Theoretical value for any d (except zero) is - for a random series with independent observations equal to zero. Therefore, if we receive for an observed not cyclical time series r, near zero it means that our time series is random, (assuming that a sample size is not very small). If r_1 is close to -1 or to 1, observations of our time series are not independent.

We may employ approximated rule that value of r_d exceeding $2/\sqrt{N}$ shows that observations of time series do not represent a purely random process. Probability that our statement is untrue is 0.05.

Highly autocorrelated time series are utilized for the forecasting purpose. High autocorrelation shows that X_t (actual) observations are strongly determined by previous time series values X_{t-d}. It means that past time series observations could be utilized as explaining variable for dependent variable X_t. Since autocorrelation is the measure of linear association between X_t and X_{t-d} our model should be linear regression model.

$$X_t = a + b\, X_{t-d} + e_t \qquad (3.2)$$

$$t = d+1,\ d+2,\ \ldots,N$$

The formulas for calculating b and a are the same as for the linear regression between Y and X variables (see Chapter 1, Section1.3), X_t and X_{t-d} variables being represented in this formulas by Y and X variables.

Having estimated a and b parameters forecasted value may be calculated as follows:

$$X_{N+1} = a + b\, X_{N+1-d} \tag{3.3}$$

If for example d=1 formula (3.3) becomes

$$X_{N+1} = a + bX_N \tag{3.4}$$

Since for a straight line relation between two variables coefficient of determination (R^2) is equal to squared correlation coefficient between Y and X, squared autocorrelation coefficient r_d^2 shows what part of X_t series variance could be explained by its delayed observations.
For example if $r_d = 0.96$ $R^2 = r_d^2 = 0.92$.

Apart from the straight line (3.2) sometimes for forecasting purposes multiple regression relation is used, e.g. equation with two explaining variables.

$$X_t = a + bX_{t-1} + cX_{t-2} + e_t \tag{3.5}$$

and non linear equations, e.g. exponential

$$X_t = ab^{X_{t-d}}\, e_t \tag{3.6}$$

power

$$X_t = a\, X_{t-d}^{b}\, e_t \tag{3.7}$$

or logarithmic

$$X_t = a + b \log X_{t-d} + e_t \tag{3.8}$$

Application of multiple regression equation is justified when actual value of X_t is determined by many preceding observations. Application of (3.6) - (3.8) relations is justified when association between X_t and X_{t-d} is not such straightforward as in equation (3.2). For (3.6) we expect that there should be high correlation between log X_t and X_{t-d} rather than between original observations of X_t. For (3.7) there should be high correlation between log X_t and log X_{t-d} and for (3.8) between X_t and log X_{t-d}, while correlation between original observations

may be rather small. Models like (3.2) and (3.5) -(3.8) are called autoregressive models.

Mentioned above autocorrelations may serve as a statistical criterion for the choice of the appriopriate forecasting model. Their squared values in the cases of exponential and power equations show determination of log X_t variable variances by a straight line regressions of X_{t-d} and log X_{t-d} variables respectively.

In the case of logarithmic relation (3.8), r_d^2 shows determination of original X_t variable by log X_{t-d} variable. Therefore determination coefficients for the (3.6) and (3.7) models should be calculated as

$$R^2 = 1 - \frac{Var(\hat{e}_t)}{Var\ (X_t)} \qquad (3.9)$$

where

$Var(\hat{e}_t)$ stands for variance of residuals

$$\hat{e}_t = X_t - \hat{X}_t$$

$Var\ (X_t)$ stands for variance of X_t.

3.3. FORECASTING TIME SERIES- CROSS SECTION DATA USING SIMPLE AUTOREGRESSIVE MODEL

Time series-cross-section data are important class of statistical data, keeping information about variability in time and between units of the sample (regions, groups, classes, individuals, etc.).

Construction and estimation of the proper econometric models enables to reveal reasons for this two kinds of variability; changes in time and differentiation of units constituting sample.

Our objective is to present a method of forecasting on the base of time series-cross section data i.e. a set of 14 time series one for each Region in Ethiopia.

Example 1

Table 1 contains data on Area under Teff in 14 provinces of Ethiopia in five year period from 1974/75 to 1978/79

Table 1

Area under Teff by regions

Region	Area in year				
	1974/75	1975/76	1976/77	1977/78	1978/79
Arssi	24.5	27.7	24.1	25.2	29.0
Bale	8.2	8.6	7.4	5.5	8.2
Eritrea	22.4	24.9	27.1	28.1	28.7
Gamo-Gofa	13.5	14.6	11.8	13.0	14.6
Gojam	213.7	261.5	249.3	254.2	243.3
Gonder	201.5	234.1	216.6	201.3	205.9
Hararge	6.2	5.3	3.1	2.7	2.6
Illubabor	44.1	53.8	53.9	43.4	47.4
Keffa	80.4	97.5	87.8	104.8	116.9
Shoa	297.1	344.7	314.6	306.5	345.7
Sidamo	15.3	19.5	22.7	11.4	13.3
Tigrai	59.4	70.4	65.7	64.4	66.8
Wellega	152.9	184.4	180.2	158.6	177.6
Wollo	49.5	56.2	46.4	58.2	64.7
TOTAL	1188.7	1403.2	1310.7	1279.3	1365.2

Our objective is to forecast area under teff by regions for 1979/80. We are going to use autoregressive model with a set of one year delayed observations as explaining variable:

$$x_t = \varphi(x_{t-1}, e_t) \qquad (3.10)$$

The selection of the most proper function among a straight line, exponential and power functions will be performed on the base of a priori calculated determination coefficient $R^2 = r_1^2$ taking correlation between original X_t and X_{t-1}, between log X_t and X_{t-1} and between log X_t, log X_{t-1} respectively. We have following coefficients of determination:

for the straight line	0.98
for the exponential	0.74
for the power	0.98

Since area under teff in 1974/75 is quite different from the data of the remaining years (compare TOTAL in Table 1) we should try if the resignation of data from this year would improve the coefficient of determination. For four years data and for a straight line it equals 0.99. Therefore, we have decided to base our forecasting procedure on data from 1975/76 to 1978/79 applying a straight line:

$$X_{tR} = a+bx_{t-1R} + e_{tR} \qquad (3.11)$$

Table 2 shows the arrangement of data from Table 1 in squence:

dependent variable X_t,

explaining variable X_{t-1}

Table 2

Dependent variable X_{tR} and explaining variable X_{t-1R}(part)

Region	t (year)	X_{tR}	X_{t-1R}
Arssi	1976/77	24.1	27.7
	77/78	25.2	24.1
	78/79	29.0	25.2
Bale	1976/77	7.4	8.6
	77/78	5.5	7.4
	78/79	8.2	5.5
Eritrea	1976/77	27.1	24.9
	77/78	28.1	27.1
	78/79	28.7	28.1
et.c.			

As one can see in Table 2 each region is represented by three observations. Dependent variable observations are data from 1976/77, 1977/78 and 1978/79 years, explaining variable observations are data from 1975/76, 1976/77 and 1977/78 respectively.

Analysing the data for all regions, we got a big sample giving the possibility of taking the rule of area changes in all country into account. Attempting to forecast area for each region separately, for example by fitting to each regional time series a simple trend function, we would lose information on the trend at the country level. On the other hand model (3.11) does not include variable explaining variability specific for each region. This information may be recovered. The method will be presented below.

The straight line relation between X_{tR} and X_{t-1R} was estimated by least squares method:

$$X_{tR} = 0.15 + 1.01\, X_{t-1R} \qquad (3.12)$$

Theoretical values and forecast for 1979/80 calculated on the base equation (3.12) are presented in Table 3. Residuals $X_{tR} - X_{tR}$ are shown in Table 4.

Table 3

Theoretical values $\hat{X}_t$ and forecast $\hat{X}_{79/80}$

Region	Theoretical values			forecast for
	1976/77	1977/78	1978/79	1979/80
Arssi	28.1	24.5	25.6	29.4
Bale	8.8	7.6	5.7	8.4
Eritrea	25.3	27.5	28.5	29.1
Gamo-Gofa	14.9	12.1	13.3	14.9
Gojam	263.8	251.5	256.4	245.4
Gonder	236.2	218.5	203.1	207.7
Hararge	5.5	3.3	2.9	2.8
Illubabor	54.4	54.5	45.9	47.9
Keffa	98.4	88.7	105.8	118.0
Shoa	347.7	317.3	309.1	347.7
Sidamo	19.8	23.4	11.7	13.6
Tigrai	71.1	66.4	65.1	67.5
Wellega	186.1	181.8	160.0	179.2
Wollo	56.8	46.9	58.8	65.4
TOTAL	1416.9	1324.0	1291.9	1378.0

Table 4

Residuals $X_t - \hat{X}_t = \hat{e}_t$

1976/77	1977/78	1978/79	Average $\bar{e}_R$	Corrected forecast
-4.0	0.7	3.4	0.0	29.4
1.4	-2.1	2.5	0.6	9.0
1.8	0.6	0.2	0.9	30.0
-3.1	0.9	1.3	-0.3	14.6
-14.5	2.7	-13.1	-8.3	237.1
-19.6	-17.2	2.8	-11.3	196.4
- 2.4	- 0.6	- 0.3	- 1.1	1.7
- 0.5	- 5.1	1.5	- 2.7	45.2
-10.6	16.1	11.1	5.5	123.5
-33.1	-10.8	36.6	- 2.4	346.3
2.9	-11.6	1.6	- 2.4	11.2
- 5.4	- 2.0	1.7	- 1.9	65.6
- 5.9	-23.2	17.6	- 3.8	175.4
-10.4	11.3	5.9	2.3	67.7
-103.4	-40.3	72.8	-24.9	1353.1

Function (3.12) estimates relations between X_t and X_{t-1} variables only generally as average relations for all country. This relations are defined by a and b parameters. They can give a good approximation for all country but they explain regional differenciation for relation between X_t and X_{t-1} only partly. Regional discrepancy from the relation (3.12) is included, into error term of model (3.12), i.e. into $\hat{e}_t$. We may check that for some regions residuals $\hat{e}_t$ are of the same sign. It means that our forecasted values for regions should be shifted down or up depending on the sign of the residuals. This can be done by adding to the forecasted values arithmetic averages of residuals calculated for every region. These averages are shown in the Column 4 of Table 3. Column 5 comprises corrected forecast calculated as follows

$$X^*_{79/80} = \hat{X}_{79/80} + \bar{e}_R$$

It should be noted that correcting values $\bar{e}_R$ minimalise sum of squares

$$D_R = \sum \left[X_{tR} - (\hat{X}_{tR} + d_R)\right]^2 \qquad (3.13)$$

where d_R stands for unknown correcting value. According to the rule for finding of minimal or maximal value of the function we receive:

$$\frac{\partial D_R}{\partial d_R} = 2\left(\sum X_{tR} - \sum d_R - \sum \hat{X}_{tR}\right) = 0 \qquad (3.14)$$

Solving equation (3.14) we receive

$$d_R = \frac{\sum (X_{tR} - \hat{X}_{tR})}{n} = \frac{\sum \hat{e}_{tR}}{n} = \bar{e}_R \qquad (3.15)$$

Correcting coefficient d_R minimizing the sum of squares

$$D_R = \sum (X_{tR} - d_R \hat{X}_{tR})^2 \qquad (3.16)$$

should be applied for models with multiplicative error term such as exponential

$$X_t = a\, b^{X_{t-1}}\, e_t$$

or power

$$X_t = a\, X_{t-1}^{b}\, e_t$$

This coefficient can be found as the solution of equation

$$\frac{\partial D_R}{\partial d_R} = 2\sum (X_{tR} - d_R \hat{X}_{tR})\, \hat{X}_{tR} \qquad (3.17)$$

and therefore

$$d_R = \frac{\sum X_{tR}\, \hat{X}_{tR}}{\sum X_{tR}^2} = 0 \qquad (3.18)$$

where R means that we should take data for each region separately. Corrected forecast is calculated as follows:

$$x^*_{79/80} = \hat{x}_{79/80} \cdot d_R$$

Coefficient d_R can be calculated also for the straight line relation but its application give almost the same result as application of the simple averages (3.15). .

Forecast of the area under crops, production and yields of major crops for total Ethiopia and by regions for 1979/80 and 1980/81 were performed and published in 1979(see [22]).

This forecast is based on cross-section-time series data and utilizes power autoregressive function (3.7).

3.4. PROJECTION OF TRENDS APPLIED TO CROSS-SECTION-TIME SERIES DATA

Having at our disposal data on the development of examined phenomena over time at the regional level we want to forecast its regional values.

The simplest method consists of estimation and extrapolation of the trend adjusted to each regional time series separately. However, this straightforward method sometimes gives unsatisfactory results since extrapolated values of regional trend may be inconsistent with tendency for all country. The reason for it is that attempting to consider each region separately one does not take into account interrelations between regional time series.

Internally consistent estimates of regional trends result from model adjusted to cross-section-time series data. Such model should consist of trend at the country level and trends for regional effects. Country level trend is something as trend for non existing "average region". Regional effects show regional differences from this average trend. One possible version of such model may be presented as follows

$$X_{tR} = \varphi(t) + \sum_{r=1}^{14} \gamma_r(t)\delta \qquad (3.19)$$

where

$\varphi(t)$ stands for the country level trend

R is number of region R= 1,2,..., 14

γ_r are functions approximating trends of regional effects

δ is dummy variable

$$\delta = \begin{cases} 1 & \text{if } r = R \\ 0 & \text{if } r \neq R \end{cases}$$

Model (3.19) consists of one function for country level trend and one function for each region, totally 15 functions. If each of them is thought as a simple two parameter function it means that one should estimate multiple regression equation with 30 parameter what requires use of electronic computer program.

However, this forecasting model may be realized approximately without use of electronic computer, through application of programmable desk calculator is indicated.
To present this method we should notice that observation from year t and region r may be considered as a sum or product of (A_t) average level in year t, (R_{tr}) regional effect and random error e_{tr}: i.e. additive model

$$X_{tr} = A_t + R_{tr} + e_t \qquad (3.20)$$

or multiplicative model

$$Y_{tr} = A_t \cdot R'_{tr} \cdot e'_t \qquad (3.21)$$

$$t = 1,2,\ldots, T$$

where

$$A_t = \left(\sum_{r=1}^{14} X_{tr} \right)/14. \qquad (3.22)$$

Because $14\ A_t = \sum X_{tr}$, then R_{tr} and R'_{tr} should fulfil following conditions

$$\sum R_{tr} = 0 \qquad \sum R'_{tr} = 14. \qquad (3.23)$$

Our task is estimation and projection of the trends for average levels A_t and for regional effects R_{tr} (R'_{tr}).
It may be done by means step by step procedure similar to that (presented in Chapter II) for changing seasonality.
This method can be performed as follows.

1. Estimation of the trend of average level. Trend may be approximated by means of simple function φ (straight line, exponential, logarithmic or power) adjusted to series of averages

$$A_t = \varphi(t, e_t) \tag{3.24}$$

2. Calculation of theoretical average level on the basis of function (3.24)

$$\hat{A}_t = \varphi(t) \tag{3.25}$$

and of average level projected for year T + 1

$$\hat{A}_{T+1} = \varphi(T + 1) \tag{3.26}$$

3. Adjustment of trends to deviations from theoretical average levels A_t, for each region separately, i.e to differences

$$D_{tr} = X_{tr} - \hat{A}_t \tag{3.27}$$

or to ratios

$$D_{tr} = X_{tr}/\hat{A}_{tr} \tag{3.28}$$

These trends may be approximated by means of the simple mathematical functions γ_r (straight line, exponential, logarithmic, power), i.e.

$$D_{tr} = \gamma_r (t, e) \tag{3.29}$$

4. Calculation of the estimated regional effects as theoretical values from the function (3.29)

$$R_{tr} = \gamma_r (t) \tag{3.30}$$

5. R_{tr} (R'_{tr}) for each year should be adjusted to fulfil condition (3.23). For additive regional effects it may be done by subtraction from each R_{tr} corresponding annual average of R_{tr}s. For multiplicative regional effects each of the R'_{tr} should be divided by corresponding annual average of R'_{tr}s (compare adjustment of seasonal deviations and coefficients, Chapter II, Section 2.5.6).

6. Projection of the regional effects to year T + 1.

$$R_{T+1,r} = \gamma_r(T+1) \tag{3.31}$$

7. Calculation of the forecasted values for regions

$$\hat{X}_{T+1,r} = \hat{A}_{T+1} + R_{T+1,r} \tag{3.32}$$

or $$\hat{X}_{T+1,r} = \hat{A}_{T+1} \cdot R_{T+1,r} \tag{3.33}$$

8. Calculation of the residuals

$$\hat{e}_{tr} = X_{tr} - \hat{X}_{tr}, \tag{3.34}$$

standard error

$$S_e = \sqrt{\sum_{t,r} e_t^2 /(T \cdot 14)} \tag{3.35}$$

and coefficient of determination

$$R^2 = 1 - \left(\sum_{t,r} e_{tr}^2 \right) / \sum_{t,r} x_{tr}^2 \tag{3.36}$$

where $x_{tr} = X_{tr} - \bar{X}$.

Applying this method special attention should be paid to possible occurence of extrapolated values going far outside limits of admissible forecasted values.

Example

Having data on the Yield of Teff in regions in 1974/75 to 1978/79 we want to forecast yield of teff in 1979/80.

Table 5

Yield of Teff in 1974/75 - 1978/79 by regions

Region	Yield in year				
	1974/75	1975/76	1976/77	1977/78	1978/79
Arssi	8.0	7.9	13.5	13.2	11.3
Bale	5.6	8.4	5.9	6.6	6.5
Eritrea	7.0	7.2	7.1	7.4	7.4
Gamo Gofa	4.6	6.4	4.5	4.4	4.3
Gojam	6.8	4.6	6.9	7.9	7.4
Gondar	6.1	4.2	6.1	9.6	10.0
Hararge	4.3	3.4	5.8	6.1	7.0
Ellubabor	5.4	6.5	5.0	5.9	6.9
Kefa	7.0	9.4	6.9	7.7	7.8
Shoa	7.3	9.0	8.8	7.6	7.8
Sidamo	7.8	12.5	9.●	9.3	9.5
Tigrai	9.1	7.2	7.3	7.5	7.5
Wellega	6.6	7.8	7.6	5.0	4.7
Wollo	9.2	7.2	9.5	11.2	10.4
Average Levels	6.8	7.3	7.4	7.8	7.8
Trend	6.8	7.2	7.5	7.7	7.8

Because time series of average levels consists barely of 5 observations the only criterion for the choice of function approximating trend is goodness of fit. We received following coefficient of determination.

- straight line $R^2 = 0.889$
- exponential $R^2 = 0.882$
- logarithmic $R^2 = 0.958$ (-)
- power $R^2 = 0.960$

Our choice is for the power curve:

$$A_t = 95.07\ t^{0.089} \qquad (3.37)$$

Trend values calculated on the basis of (3.37) are shown as the last row of table 5. Trend extrapolated for 1979/80 is 8.0

Next step is estimation of the regional effects. We arbitraly decided to apply multiplicative model (3.21)

Ratios $X_{tr} / \hat{A}_t$ are shown in Table 6.

Table 6

Ratios $D'_{tr} = X_{tr} / \hat{A}_t$

Region		1974/75	1975/76	1976/77	1977/78	1978/79
Arssi	P	1.177	1.097	1.800	1.714	1.449
Bale	A	0.824	1.167	0.786	0.857	0.833
Eritrea	P	1.029	1.000	0.947	0.961	0.949
Gomo Gofa	E	0.676	0.889	0.600	0.571	0.551
Gojam	A	1.000	0.639	0.920	1.026	0.949
Gondar	L	0.897	0.583	0.813	1.246	1.282
Hararge	L	0.632	0.472	0.773	0.792	0.897
Illubabor	A	0.794	0.903	0.667	0.766	0.885
Kefa	A	1.029	1.306	0.920	1.000	1.000
Shoa	E	1.074	1.250	1.173	0.987	1.000
Sidamo	A	1.147	1.736	1.200	1.208	1.218
Tigrai	P	1.338	1.000	0.973	0.974	0.962
Wellega	E	0.971	1.083	1.013	0.649	0.603
Wollo	L	1.353	1.000	1.267	1.455	1.333

econd column of table 6 shows what function was choosen on the basis of goodness of fit criterion to represent trend of D'_{tr} ratios from corresponding row of table. P - denotes power function, E - exponential, L - straight line and A denotes constant calculated as arithmetic means. For example the trend line fitted to the ratios for Arssi region is described by power function

$$D'_{t1} = 1.137t^{0.231} \qquad R^2 = 0.45 \qquad (3.38)$$

For Gamo Gofa was choosen exponential function

$$D'_{t4} = 0.835\, e^{-0.085\, t} \qquad R^2 = 0.48 \qquad (3.38)$$

(e is basis of natural logarithm).
For Harrar the best fitted function was the straight line

$$D'_{t7} = 0.458 + 0.085t \quad R^2 = 0.67 \qquad (3.39)$$

et.c.

Theoretical values calculated on the basis of adjusted functions we shall call Preliminary Coefficients of Regional Effects. They should be arranged in table, similar to Table 6, one column for a year. To receive final coefficients we should adjust data from each year to total 14, according to condition (3.23). It was done by dividing numbers in column by their arithmetic mean. Adjusted data, i.e. Final coefficients of Regional Effects are presented in Table 7.
Multiplying coefficients from Table 7 by corresponding theoretical Average Levels A_t we received theoretical values of the yield of Teff, X_{tr} (Table 8).
Forecasted values for 1979/80 are shown in Table 9, where they are compared with forecast received on the basis of autoregressive model published in July 1979 [22] .

Table 7.
Final Coefficients of Regional Effects R'_{tr} calculated on the basis of functions fitted to rows of Table 6 and adjusted by columns to total 14

Region	1974/75	1975/76	1976/77	1977/78	1978/79	1979/80 (projected)
Arssi	1.143	1.346	1.472	1.562	1.632	1.686
Bale	0.898	0.901	0.897	0.891	0.884	0.876
Eritrea	1.033	0.999	0.973	0.953	0.933	0.916
Gamo-Gofa	0.771	0.711	0.650	0.594	0.540	0.491
Gojjam	0.912	0.915	0.911	0.905	0.898	0.890
Gonder	0.681	0.828	0.968	1.105	1.238	1.367
Harrarge	0.546	0.634	0.716	0.796	0.874	0.949
Illubabor	0.807	0.810	0.806	0.801	0.795	0.788
Keffa	1.057	1.060	1.056	1.049	1.040	1.031
Shoa	1.184	1.144	1.097	1.050	1.003	0.956
Sidamo	1.309	1.313	1.308	1.299	1.289	1.277
Tigrai	1.264	1.106	1.016	0.954	0.906	0.865
Wellega	1.131	0.981	0.843	0.723	0.620	0.530
Wollo	1.204	1.251	1.288	1.320	1.351	1.379

Table 8
Theoretical values $\hat{X}_{tr} = \hat{A}_t \cdot R_{tr}$

Region	1974/75	1975/76	1976/77	1977/78	1978/79
Arssi	7.8	9.7	11.0	12.0	12.[illegible]
Bale	6.1	6.5	6.7	6.9	6.9
Eritrea	7.0	7.2	7.3	7.3	7.3
Gamo-Gofa	5.2	5.1	4.9	4.6	4.2
Gojjam	6.2	6.6	6.8	7.0	7.0
Gonder	4.6	6.0	7.2	8.5	9.7
Hararge	3.7	4.5	5.4	6.1	6.8
Illubabor	5.5	5.8	6.0	6.2	6.2
Keffa	7.2	7.7	7.9	8.1	8.1
Shoa	8.1	8.3	8.2	8.1	7.8
Sidamo	9.0	9.5	9.8	10.0	10.0
Tigrai	8.6	8.0	7.6	7.3	7.1
Wellega	7.7	7.1	6.3	5.6	4.8
Wollo	8.3	9.0	9.6	10.2	10.5
Average Levels	6.8	7.2	7.5	7.7	7.8

Standard error of estimation S_e may be used as a measure showing possible average error of forecast. To estimate it we should calculate sum of squares of differences

$\hat{e}_{tr} = X_{tr} - \hat{X}_{tr}$. It is easy to check that for our example $\sum \hat{e}^2_{tr} = 64.0$

Since number of observations on our data is 70 and number of estimated parameters is 2 for trend of average levels plus 23 for trends of regional effects, to calculate S_e we should divide 64 by 70-25= 45. Therefore,

$$S_e = \sqrt{64/45} = 1.2 \text{ qt/hectare.}$$

Coefficient of determination is

$$R^2 = 1 - 64/302 = 0.788.$$

Standard error of the forecast based on autoregressive model (presented in Table 9) is 2.4 qu/hectare.

Table 9

Forecasted values for 1979/80 estimated on the base of projection of trends and on the base of autoregressive model (see [22] , p.51-64)

Region	Forecast based on	
	trends	autoregr.
Arssi	13.5	11.9
Bale	7.0	6.3
Eritrea	7.3	7.4
Gamo-Gofa	3.9	4.0
Gojjam	7.1	7.5
Gonder	10.9	10.8
Hararge	7.6	7.9
Illubabor	6.3	7.1
Keffa	8.2	7.5
Shoa	7.6	7.7
Sidamo	10.2	9.6
Tigrai	6.9	7.1
Wellega	4.2	4.3
Wollo	11.0	10.6
Standard error	1.2	2.4

3.5. EXPONENTIAL SMOOTHING FORECASTING MODELS

3.5.1. Description of the method

Exponential smoothing models attempt to estimate theoretical smoothed time series observations as a weighted average of its actual and all previous observations with weights decreasing exponentially so that remote observations receive smaller weights than more actual ones.

The idea of exponential smoothing can be presented in the following way. Let us $\hat{X}_t$ denotes smoothed time series observation corresponding to real one X_t. It can be expressed as a weighted average of actual real observation and previous smoothed observation, i.e.

$$\hat{X}_t = a\, X_t + (1-a)\hat{X}_{t-1} \tag{3.40}$$

substituting for $\hat{X}_{t-1}$, $\hat{X}_{t-2}$, ..., $\hat{X}_{t-j}$, ... into (3.40) we get

$$\hat{X}_t = a \sum_{j=0} (1-a)^j X_{t-j} \tag{3.41}$$

The latest available smoothed value X_N is employed to forecast future value of the series, i.e.

$$\hat{X}_{N+h} = \hat{X}_N \qquad h = 1,2,... \tag{3.42}$$

where $\hat{X}_{N+h}$ denotes forecasted value for h units of time ahead. Parameter a may be choosen to minimize the sum of squared errors between realizations and forecasts made for past units of time for whose time series observations are available.

In practice this simple formulation is rarely employed and several modifications have been developed. One of the widely applied approaches due to Holt [9] and Winters [18]

in its the simplest version for nonseasonal time series, assumes that time series observation at time t is composed of local mean level M_t and random error e_t, which is assumed to be of constant variance. Local mean at time t differs from local mean at time t-1 by value of local trend T_t.

Local mean at time t is estimated as weighted average of time series observation at time t and forecasted value for time t. Forecast for time t is simply sum of local mean and local trend at time t-1, i.e.

$$\hat{X}_t = M_{t-1} + T_{t-1} \tag{3.43}$$

Therefore, estimation of local mean can be presented as follows

$$M_t = a\,X_t + (1-a)\,(M_{t-1} + T_{t-1}) \tag{3.44}$$

or

$$M_t = a\,X_t + (1-a)\,\hat{X}_t \tag{3.45}$$

where $o < a < 1$

Trend at time t is defined as difference between (true) mean levels at time t and (t-1). Estimator of trend is following

$$T_t = b(M_t - M_{t-1}) + (1-b)T_{t-1} \tag{3.46}$$

$o < b < 1$

The rule for forecasting one unit of time ahead is given in (3.43). Forecast for N+h time, when X_N is last available time series data is made on the basis of the last estimated mean level and trend

$$\hat{X}_{N+h} = M_N + hT_N \tag{3.47}$$

Therefore, forecasting model consists of two updating formulas (3.44) and (3.46), applied as new time series observation is available, and of formulation of forecasting rule

(3.47). It requires to determine suitable values for the smoothing constants a and b and starting values for M_t and T_t.

Constants may be choosen on the empirical ground by means of trial forecasts made one time period previously of the known observations X_t. That values a and b which best "foresee" the known observation (generally in term of average squarred forecast error) are then employed as forecasting constants.

Sometimes suggestion for the choice of the most proper parameters results from the character of data. We tried to find the best parameters for time series of ratios D'_{tr} from Table 6 The best ability to forecast, measured in terms of the sum of squared differences between forecast and real value, have following models:

1) a= 0.3, b= 0.3 for Arssi, where ratios are highly irregular but show slight increasing trend,
2) a= 0.8, b= 0.3 for Bale where ratios are irregular with rather slight trend,
3) a= 0.9, b= 0.4 for Gamo-Gofa where data show distinct decreasing trend without big irregular variability,
4) 0.9 and 0.2 for Gonder where distinct trend is evident,
5) 0.1 and 0.1 for Shoa where ratios are slightly decreasing,
6) for Wollo where data are highly irregular nearly the same accuracy is with a= 0.3, b= 0.3 as with a= 0.9, b= 0.3

It should be noted that small value of b parameter gives bigger possibility for demonstration of steady increase or decrease of time series observations, while parameter a close to unity makes model more sensitive for new trends introduced with

actual observation. For all examined regions the best proved small parameters b, stressed steady trend, while parameter a for some regions is high and low for others, stressing in the former case modification of mean level and in the latter one its stability.

3.5.2. Application to agricultural forecasting

To begin performance of the forecasting procedure we should establish starting value of a local level and local trend. It seems that for nonseasonal data it could be simply first time series observation for local mean and zero for local trend. The procedure of forecasting is shown in Table 9, where we attempted to forecast Yield of Teff in Shoa region for 1979/80. Data (taken from Table 5) are shown in the first column of table. Further columns shows consecutive steps of procedure as a new time series observation becomes available. It follows from the procedure that to perform forecasting as new observation X_{t+1} becomes available one needs to use only this observation and last estimations of level and trend. Beginning with step 2 values of M were calculated on the basis of formula (3.45) while forecasted value ($\hat{X}_t$) for step 2 is simply repeated time series observation from previous year ($\hat{X}_2=X_1$) for further steps $\hat{X}_t=M_{t-1} + T_{t-1}$. Therefore, yield of teff forecasted for 1979/80 is calculated as

$\hat{X}_{79/80}= M_{78/79} + T_{78/79}$. Table 9 shows attempt to make forecast with a= 0.8, b= 0.2. But we should check if another parameters do not produce better criterion of accuracy. Searching for the best set of constants is shown in Table 10.

Table 9

Steps of forecasting procedure with constants:

a= 0.8, b= 0.2

X_t	Names of calculated value	Steps of forecasting procedure					
		I	II	III	IV	V	VI
7.3	X_1	—					
	M_1	7.3	7.3				
	T_1	0	0				
9.0	X_2		7.3	7.3			
	M_2		8.7	8.7			
	T_2		0.3	0.3			
8.8	X_3			9.0	9.0		
	M_3			8.8	8.8		
	T_3			0.3	0.3		
7.6	X_4				9.1	9.1	
	M_4				7.9	7.9	
	T_4				0.1	0.1	
7.8	X_5					8.0	8.0
	M_5					7.8	7.8
	T_5					0.1	0.1
Forecast	X_6						7.9

$$M_t = \alpha X_t + (1-\alpha) X_t$$

$$M_2 = \alpha X_2 + (1-\alpha) X_{2}$$

Table 10
Forecasting of Yield of Teff in Shoa, for 1979/80.

X_t	Name of calculated value	Estimations with parameters: 0.8/0.8	0.9/0.1	0.3/0.8	0.3/0.3	0.1/0.1
7.3	$\hat{X}_1$	-	-	-	-	-
	M_1	7.3	7.3	7.3	7.3	7.3
	T_1	0	0	0	0	0
9.0	$\hat{X}_2$	7.3	7.3	7.3	7.3	7.3
	M_2	8.7	8.8	7.8	7.8	7.5
	T_2	1.1	0.2	0.4	0.2	0.0
8.8	$\hat{X}_3$	9.8	9.0	8.2	8.0	7.5
	M_3	9.0	8.8	8.4	8.2	7.6
	T_3	0.5	0.1	0.5	0.2	0.0
	$(X_3-\hat{X}_3)^2$	1.00	0.04	0.36	0.64	1.69
7.6	$\hat{X}_4$	9.5	8.9	8.9	8.4	7.6
	M_4	8.0	7.7	8.5	8.2	7.6
	T_4	-0.7	0.0	0.2	0.1	0.0
	$(X_4-\hat{X}_4)^2$	3.61	1.69	1.69	0.64	0.00
7.8	$\hat{X}_5$	7.3	7.7	8.7	8.3	7.6
	M_5	7.7	7.8	8.4	8.2	7.6
	T_5	-0.4	0.0	0.0	0.1	0.0
	$(X_5-\hat{X}_5)^2$	0.25	0.01	0.81	0.25	0.04
Forecast	-	7.3	7.8	8.4	8.2	7.6

First parameters a= 0.8, b= 0.2 were checked, producing K_F = 2.29(compare Table 9).

$$K_F=\sum (X_t-\hat{X}_t)2 \qquad (3.48)$$

It should be noticed that first forecasted value is for third year, $\hat{X}_3$, because $\hat{X}_2$ is simply repeated X_1. First three steps of forecasting are usually running-in-procedure steps. Then for calculation of accuracy measure K_F are taken $(X_t-\hat{X}_t)^2$ beginning from t=4

Next we increased parameter b to 0.8 a being not changed (a=0.8) getting K_F= 3.61. Since increasing of b resulted in worse forecast we came back to low b values, getting for a= 0.9, b= 0.3 K_F = 2.89 and for a= 0.9, b= 0.1 K_F= 1.70. So far this last result proved the best. But we next checked small value of a and big value of b, and small both a and b. We received following values of K_F:

for a= 0.3 b= 0.8 K_F= 1.70
for a= 0.3, b= 0.3 K_F= 0.89
for a= 0.1, b= 0.1 K_F= 0.04

This last set of parameters was choosen as the best one. It produces forecasted value 7.6. In the similar way we received forecasts for Arssi, Bale, Gamo-Gofa, Gonder and Wollo. They are shown in Table 11, where they are compared with earlier forecasts made on the base of power autoregressive function (see [22]) and on the bases of extrapolated trends.

Table 11

Forecasts of Yield of Teff for 1979/80 from exponential smoothing, autoregressive model and projection of trends, for choosen regions.

Region	Forecast from			Parameters
	Exp. smooth.	Autoregr.	Trend	for exp. smooth.
Arssi	13.1	11.9	13.5	0.5,0.3
Bale	6.5	6.3	7.0	0.8,0.3
Gamo-Gofa	4.1	4.0	3.9	0.9,0.4
Gonder	10.8	10.8	10.9	0.9,0.2
Shoa	7.6	7.7	7.6	0.1,0.1
Wollo	10.8	10.6	11.0	0.9,0.3
Standard error S_e	1.1	2.4	1.2	

For comparison with accuracy of another methods standard error S_e was calculated on the base of sum of squarred differences between smoothed and original observations, i.e.

$$S_e = \sqrt{\sum (X_t - M_t)^2/(30-18)} \qquad (3.49)$$

where (30-18) shows that 12 parameters (two for each region) were adjusted and smoothed value for the first year was taken as equal to original observation.

The better seems to be another measure of accuracy showing standard deviation of differences between original observations and their forecast. It should be calculated as follows

$$S_F = K_F/N \qquad (3.50)$$

where K_F is determined in formula (3.48),
N is number of differences $X_t - X_t$ taken to calculation of K_F

For calculation of S_F three first years should be dropped. For the exponential smoothing forecast presented in Table 11, S_F= 1.45 i.e. average error of our forecast is 1.45 qu/hectare.

Since we had data only for five year period, S_F was calculated on the base of results for the last two years only. Therefore, in our example N= 12(six regions times 2 years).

We attempted to apply 2 parameter Holt Winters model to forecast yield of teff by regions considering the data on yield in a similar way as in section 3.4 , i.e. observation X_{tR} being composed of mean level in year t, X_t, regional deviation from this mean level R_t and error term e_{tR}. As first step we forecasted for 1979/80 mean level using data on mean levels from Table 5 . We received the same forecasted value, 8.0 as from power function applied in Section 3.4. Next, similarly like it was done in the method with extrapolated trends, ratios original/smoothed mean level were calculated and for each region separately forecast of ratios was performed. Results for 6 regions are shown below. Accuracy measures

$$S_e = 1.2, \quad K_F = 1.5$$

are similar to these received from the simpler, more straightforward approach based on original data for regions. Forecasted values received from exponential smoothing with decomposition of observations:

Arssi	12.9
Bale	6.5
Gamo-Gofa	4.0
Gonder	10.7
Shoa	7.8
Wollo	10.3

3.5.3. Exponential smoothing forecast of seasonal data

For data showing distinct seasonal variation Holt-Winters forecasting method requires to apply additional third parameter, C. The model may be presented as follows (compare [4])

$$M_t = aX_t/S_{t-p} + (1-a)(M_{t-1} + T_{t-1}) \quad (3.51)$$

$$S_t = cX_t/M_t + (1-c)\, S_{t-p} \quad (3.52)$$

$$T_t = b(M_t - M_{t-1}) + (1-b)T_{t-1} \quad (3.53)$$

where p is period of seasonal variation, for time series with monthly observations p= 12, for quarterly time series p=4.

S_t stands for seasonal variation.

Recommended starting values (compare [4, 23] are following

a) Mean should be taken as the average observation in the first year.

b) Trend should be set to zero or to average monthly difference between the first and second years averages.

c) Seasonal factors calculated from the first years' data by comparing each observation with the overall average in the first year.

It should be kept in mind that starting values as well as parameters depend to some extent on the properties of data.

Seasonal model requires a big amount of calculation. For each time series, observation should pass through three updating formulas (3.51)-(3.53). Taking into consideration that the best set of parameters should be choosen on empirical ground it makes use computer program indispensable.

Disadvantage of the exponential smoothing procedures is that they restrict forecasting model to only one form. The forecasted value depends on the choice of parameters. Therefore, automatization of forecasting procedure is recommended, as application of computer program enables choice of the optimal constants. On the other hand in many cases,

particularly in agricultural statistics, data requires individual nonautomatic approach. Thus, solution may be automatization of the most labor-consuming calculations with possibility for analysis of results on every step of calculation.

3.6. REMARKS ON THE LIMITATIONS OF STATISTICAL FORECASTING

On account of the nature of economic and social phenomena their future value can not be predicted precisely. As a result of human activity and influence of nature this phenomena are partly deterministic partly random and are forecasted within random limits only.

Standard errors of estimation are applied to construction of limits in which true forecasted value should fall with satisfactory high probability. If random errors of forecasting function or smoothing model adjusted data are independent and normally distributed with mean value zero then true forecasted value $\hat{X}$ is situated in the partition defined by forecasted value X and standard error S_e, i.e.

$$\hat{X} - 1.96S_e \leq X \leq \hat{X} + 1.96S_e \qquad (3.54)$$

with probability 0.95 (95%). If assumptions about the error term are fulfiled relation (3.54) shows the impact of the random unpredictable part of process on the forecast.

The second reason which can cause that forecasted value may differ considerably from realization is that forecasts are made only on the base of statistical data from the past. Applied models take into account relations between past observations of the forecasted time series (autoregressiv schemes, exponential smoothing) or attempt to describe

behaviour of time series in terms of mathematical functions of time variable t (models of trend and seasonality).

In this approache forecast is made as projection into the future regularities which occurred in the past and were estimated by means of model. Therefore, this relations or/ and regularities are estimated for average conditions which occurred in the time from which statistical data were applied to model. If in the year of forecast conditions differ considerably from this average of the past realization of the process may fall outside of standard error limits.

It shall be kept in mind, however, that sometime big discrepancy between realization and forecast is important statistical information. It can inform policymaker about possibility of turning point in trend or about occurrence of situation which requires further statistical exploration. It can show also how effective are economical or administrative measures taken by policy-makers.

3.7. THE DRAFT OF MULTIPLE REGRESSION FORECASTING MODELS

3.7.1. The general description of multiple regression

Basic ideas on the simple regression model in which the dependent variable is a linear function of one independent (explanatory) variable are presented in Chapter I.

In general, economic and social phenomena are influenced by many different variables. Then in statistical analysis very often multiple regression models are more appropriate than models with one or two independent variables.

Such model may be written as

$$Y_i = b_o + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + e_i \quad (3.55)$$

$$i = 1, 2, \ldots, N$$

Non linear equations, for instance

$$Y_i = b_o X_{1i}^{b1} X_{2i}^{b2} \ldots X_{ki}^{bk} e_i \quad (3.56)$$

are often applied, but for estimation by means of least squares method they need to be transformed to linear form (3.55) by use instead of original observations their logarithms.

The least squares estimators of the parameters in (3.55) are obtained by minimizing the sum of the squared residuals

$$D = \sum e_i^2 = \sum (Y_i - b_o - b_1 X_{1i} - \ldots - b_k X_{ki})^2 \quad (3.57)$$

with respect to b_j, $j = 1, \ldots, k$.

Differentiating partially D with respect to b_j $j=0, 1, \ldots, k$ and setting the k+1 partial derivatives equal to zero we receive k+1 normal equations from which formula for estimator of parameters vector is derived. This formula in general case requires use of matrix notation. Detailed description of the least squares method applied to multiple regression equation is presented in many books on econometric methods, for example [11, 12] . Here we confine ourselves to general interpretation of parameters, assuming that estimation procedure should be performed with electronic computer using program for multiple regression analysis.

Every particular parameter b_j ($j = 1,2,\ldots,k$) from linear model (3.55) should be interpreted as amount by which dependent variable changes its level when corresponding independent variable changes its level by one unit, taking into account and excluding influence of other variables considered by model. Parameter b_o shows what would be value of dependent variable Y if all explaining variables would be at zero level.

Thus $b_3 = -0.25$ means that increase of variable X_3 by one unit cause decrease variable Y by 0.25 units, assuming that other variables do not change their level (ceteris paribus condition). Parameter B_j in model (3.56) shows that change of X_j variable q times cause the change of variable Y q^{b_j} times.

The assumption we have previously made concerning the statistical properties of the error term are the same for multiple regression equation (compare Chapter I Section 1.5). The model needs additionally one extension and one restriction.

The extension is that we now assume that every independent variable from model is treated as set of constant values. The restriction is that there should be more observations than dependent variables.

Under assumption that e_i are normally distributed testing of regression parameters significance can be applied as well as testing contribution of each explanatory variable in variability of dependent variable. As a measure of goodness of fit the coefficient of determination is defined as

$$R^2 = 1 - \sum e_i^2 / \sum y_i^2 \qquad (3.58)$$

where $y_i = Y - \bar{Y}$.

However for multiple regression more justified is so called adjusted coefficient of determination

$$R^2 = 1 - \frac{\sum e_i^2/(N-k)}{\sum y_i^2/(N-1)} \qquad (3.59)$$

which takes into consideration loss of degress of freedom connected with introduction every new variable into regression equation.

3.7.2. Possibility of use of multiple regression models to forecasting crop production and area under crops

a) Crop production

Agricultural production results from many reasons such as area under crop, meteorological conditions in different periods of plant development, use of fertilizers, occurring of pests and many others measurable and unmeasurable factors. It would be possible to build and estimate an econometric model of crop production for Ethiopia if following conditions would be fulfiled

1. We have got appropriate data concerning area, weather conditions in different periods of vegetation, use of fertilizers, et.c., for many years and at least at regional level.
3. We have got possibility to perform estimating procedure using program for electronic computer.

In the investigation performed by Statistics Section PPD, Ministry of Agriculture farmers were asked about main reasons for reported by them changes (decrease and increase) of production in 1978/79(1971E.C.) compared to last year. Results of this survey, published in [24] , show that main reasons for changes of production are weather conditions, pests and plant deseases, changes of area and use of fertilizers.
Weather conditions can be characterized by means of measurable variables such as amount of rainfall, occurence of periods with excessive rainfall, average and minimal temperatures in early vegetation period. Use of fertilizers should be expressed as number of kilograms per hectare.
Pests and plant deseases rather cannot be expressed as measurable variable and for econometric model we should apply simplified method consisting in recording of the fact of occurence. We may apply so called dummy variable, which takes value 1 if in the observed year and region pest occurred and

0 if did not. Dummy variables could be used also for the hail and flood. However it seems that pests deseases, hail and flood rather rarely effect seriously production at regional level and thus their application seems to be justified only in model based on data on small administrative units such as weredas.

One should notice that relationship between production and rainfall can be non linear. Exessive rainfall may prove destroying for production to similar extent as shortage of rain. Therefore, variables standing for rainfall could be expressed as squared deviations from normal rainfall. Influence of rainfall should be expressed by means of many variables, for example rainfalls in the period of sowing, germination, et.c. Data on temperature may be expressed as averages for different periods of vegetation . It may prove useful to introduce additional variable showing number of periods(for example weeks) in which minimal temperature was below lower limit. Our hypothetical model for crop production at regional level could be expressed as follows

$$Y_{tR} = \Psi(A_{tR}, F_{tR}, M_{1tR}, \ldots, M_{mtR}, T_{tR}, e_{tR}) \qquad (3.60)$$

where

Ψ -(unknown) function describing relation between Y_{tR} and sets of explanatory variables,

A_{tR}- area under crop in year t region R,

F_{tR}- quantity of fertilizers (kilograms per hectare),

$M_{1tR}, \ldots, M_{mtR}$- variables for meteorological factors influencing production,

T_{tR}- variable for other factors influencing production in long period such as better management, development of infrastructure, higher level of farmers education, et.c.

e_{tR}- random error of model.

Variable T_{tR} is called general trend of model and is represented by time variable (t= 1,2,...,N) or its function.

Variables on area and use of fertilizers are available at regional level. For nearly all region data on major meteorological factors influencing production can be collected .

b) Area under crops

Multiple regression model for area should include quite different explanatory variables then model for production. Change of area under crop should be considered as result of economical, social and political factors. First of all growth or fall of area under crop may be caused by changes of supply and demand for the investigated crop. Its supply in year t can be expressed as its production from the last harvest, demand can be approximated by price in period of sowing and period preceeding sowing. Social and governmental efforts to extend cultivated area could be approximatedly expressed by trend T_t similarly as in the model for production. It seems that decision about destination of area under crop depends not only on the price for this crop but also, for some crops, may result from the prices for other crops. Therefore, our hypothetical model could be expressed as

$$A_{tR}=\varphi(P_{tR}, C_{tR}, D_{1tR1}\cdots, D_{ktR1}\ T_{tR}, e_{tR}) \qquad (3.61)$$

where

A_{tR} - area under crop in year t, region R
P_{tR} - production of crop in previous year, region R
C_{tR} - price of crop in year t, region R
$D_{1tR}, \cdots, D_{ktR}$ - prices of other crops.

c) Benefits and limitations of multiple regression models

Just as we have econometric model estimated and verified it can be used for statistical and economic analysis. Estimated model for crop production shows us to what extent factors included into it as explaining variables stimulate or destimulate growth of production. It enables to set this factors in order from the most to less important stimulating/destimulating factors and even to assess numerically how they influence production. Other advantage of econometric model is that it give us possibility to estimate substitution of different factors included into it as explaining variables. For details of econometric models interpretation see [11], [12] .
But user of econometric model should be aware of its limitations. First of all numerical results are only approximation of true relations. However it is possible to assess the average size of random error of adjusted model and every estimated parameter. Second limitation is that conclusions about influence each of the explaining variables on the dependent variable are justified only in limits of their variability. It means that estimated parameters of model can show us how production changes if one of the explanatory variables changes its level, but rule of change approximated by model can be different outside the variability of this explanatory variable. It is important when one attempts to use econometric model for forecasting purpose.

d) Forecasting on the base of multiple regression model with meteorological data

If we would like to make forecast on the base of models such as (3.60) and (3.61) we should be aware of limitations which they impose. This limitations result from the model

itself and from presented above limitations imposed by general theory. First of all, forecasts based on such models can not go far into future. This limitation follows from the set of used explaining variables. Observations of this variables come from the same calendar year for which forecast is performed. The best result should give model in which meteorological variables cover all vegetation period for forecasted crop.

Then considered econometric model for production contains some kind of contradiction between its possibility to give accurate forecast and to give forecast of great usefulness. The most accurate forecast can be achieved a few month before harvest because for such period model can include variables describing weather condition and other factors influencing production in nearly all growing period. On the other hand the most useful for national economy planning are forecasts made one year earlier.

It seems reasonable that to give bigger possibilities for active and reliable forecasting Ministry of Agriculture should have for its disposal different types of models, for early prognosing (two and one year ahead) and for correcting of this early forecasting half year before and a few (3 or 2) months before the date of harvest. For each kind of forecast adequate model should be established .

3.8. BIBLOGRAPHICAL NOTE

There is extensive literature on the statistical forecasting, reflecting fast development of the theory and techniques and importance of topic for planning and management of national economy.

General theory of statistical forecasting as well as its methodology are presented in monographs of Gilchrist

[6] Granger [7] Montgomery and Johnson [15] , Wheelwright and Makridakis [17] .

Exponential smoothing was first introduced by Holt [9] and was elaborated and widely applied by Brown [2] . Other important contributors are, among others, Winters [15] who introduced the Seasonal Exponential Smoothing. Harrison 8 elaborated Harmonic Smoothing models and Trigg and Leach [16] who introduced technique of traicing signals to indicate when the characteristics of the process have changed significantly.

Autoregressive models are presented among others, in book of Malinvand [12] and paper of Newbold and Granger [23] . Autoregressive/moving-average approach, known also as the Box-Jenkins method is recently very popular among statisticians but is not easy to practical application. This method is complete and theoretically rigorous and attractive. Basical for the topic is book of Box and Jenkins[ ]containning theoretical background of method, computational formula for digital computer and providing procedures to deal with any kind of series including seasonal and nonstationary ones.
This approach is presented also in books of Granger [7]Gilchrist [6] Montgomery and Johnston [15] and in papers of Chatfield and Prothero [3] and Ma ridakis [13] . As it was mentioned above in spite of its completness and good theoretical back grounds method is not widely applied in statistical practice as it is rather complex and requires experienced expert statisticians to apply correctly.

Problems of econometric models and their use to forecasting is explored by many autors as Klein [11] , Malinvaud [12], Christ[5] , Johnson[10] .

At the end we would like to pay attention on the works of Statistics Section, PPD, Ministry of Agriculture in the field of agricultural forecasting in Ethiopia. Forecasting on the bases of field surveys is presented in their publication [19 , 20] . There had also applied statistical models of forecasting, i.e. extrapolation of trend [21] and simple autoregressive model [22] .

REFERENCES

1. Box, G.E.P. and Jenkins, G.M.(1976). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
2. Brown, R.G.(1963). Smoothing, Forecasting and Prediction. Prentice-Hall, Englewood Cliffs, New Jersey.
3. Chatfield, C. and Prothero, D.L.(1973). Box-Jenkins seasonal forecasting: Problems in a case study. Journal of the Royal Statistical Society, Series A, 136, 295-336.
4. Chafield, C. (1978). Holt-Winters forecasting procedure. Applied Statistics, Vol.I nr 3.
5. Christ, C.F. (1966). Econometric Models and Methods. John Wiley, New York.
6. Gilchrist, W.(1976). Statistical Forecasting, Wiley, London.
7. Granger, C.W.J.(1977). Forecasting Econonic Time Series. Academic Press, New York.
8. Harrison, P.J. (1965). Short-term forecasting. Applied Statistics, 14, 102-139.

9. Holt, C.C.(1957). Forecasting seasonal and trends by exponentially weighted moving averages. Carnegie Institute of Technology. Pittsburgh, Pennsylavia.

10. Johnston, J.(1972). Econometrica. Mc GrawHill, New York.

11. Klein, L.R.(1956). A Textbook of Econometrics. Evanston, Ill.

12. Malinvaud, E.(1970). Statistical Methods of Econometric. Amsterdam: North-Holland.

13. Makridakis, S.(1976). A survey of time series. International Statistical Review, 44,29-70.

14. Makridakis, S. (1978). Time series analysis and forecasting: an update and evaluation. International Statistical Review, 46,255-278.

15. Montgomery, D.C. and Johnson, L.A. (1976). Forecasting and Time Series Analysis, New York: McGraw-Hill.

16. Trigg, D.W. and Leach, D.H.(1967). Exponential smoothing with an adaptive response rate. Operational Research Quarterly, 18,53-59.

17. Wheelwright, S. and Makridaks, S.(1976). Forecasting Methods for Management: New York: John Wiley.

18. Winters, P.R.(1960). Forecasting sales by exponentially weighted moving averages. Management Science,pp. 324-342.

19. Crop production forecast for Ethiopia in 1976/77 (1969E.C.). Report, Statistics Section PPD, Ministry of Agriculture. Addis Ababa. October 1976.(mimeographed).

20. Crop production forecast for Ethiopia in 1977/78(1970E.C.) Report. Statistics Section PPD. Addis Ababa, December 1977(mimeographed).

21. Area, Production, Yield, Use of Fertilizers and Marketed Production of Major Crops. Results of the Agricultural Sample Survey 1977/78(1970E.C.) with preliminary forecast for 1978/79(1971E.C.). Statistics Section PPD. Addis Ababa, August 1978.

22. Preliminary Forecast of Area, Production and yield of Major Crops for the Whole Country and by Regions in 1979/80(1972E.C.) and 1980/81(1973E.C.). Statistics Section PPD, Addis Ababa, July 1979.

23. Newbold, P. and Granger, C.W.J.(1974). Experience with forecasting univariate time series and the combination of forecasts Journal of the Royal Statistical Society, Series A, 137,Part 2.

24. Crop Production Survey 1978/79(1971 E.C.). (Report). Statistics Section PPD. Addis Ababa June 1979.

25. Area, Production and Yield of Major Crops for the Whole Country and by Region in 1974/75-1978/79(1967E.C.-1971E.C.) (Comparisons with the Second Round Survey Results). Statistics Section PPD, Ministry of Agriculture. Addis Ababa, July 1979.