THE PROPERTIES OF RANDOM SURFACES
OF SIGNIFICANCE IN THEIR CONTACT

by

D. J. Whitehouse

Thesis submitted to the University of Leicester
for the degree of Doctor of Philosophy

February 1971

ABSTRACT

In modern engineering there is an urgent need for a deeper understanding of the nature of surface texture and its influence upon the functioning of the element of which it forms a part. Of particular importance, in this connection, is the behaviour of surfaces in stationary and sliding contact. Investigations of the contact of surfaces, on the one hand, and the evolution of methods of surface specification and characterisation, on the other, have developed more or less independently. This thesis attempts to bridge the gap between these two areas of study.

The main emphasis of the work has been upon random surfaces which are produced by a significant proportion of modern manufacturing methods. The theories used have been drawn from those employed in the study of other types of random processes. Both these theories, and the experimental evidence used to support them have been usually presented in digital form; therefore some emphasis has been placed upon the problems involved in the analysis of data presented in this form.

The theoretical analysis is concerned with the representation of a surface profile as a random signal and the significance of this for the properties of surfaces of significance in their contact. This then allows the development of a theory of the movement of a second body over such a random profile. The friction and wear of random surfaces is tackled through the analysis of results obtained from well instrumented experiments; this suggests that a stochastic approach to the tribology of random surfaces is well justified.

Finally an attempt has been made to provide a broad fundamental analysis of the generation of such surfaces. In this way it is hoped that the work provides a basis for the classification or the typology of surfaces in terms both of their functional behaviour and of the relationship of this to the details of their generation.

PREFACE

This project has been carried out partly in the Research Laboratories of the Engineering Department, University of Leicester and partly in the Metrology Research Department of Rank Precision Industries Ltd. The work has been under the supervision of Dr. J. F. Archard to whom I give my sincere thanks for encouragement, advice and helpful suggestions.

I with to thank the Directors and Management of Rank Precision Industries Ltd., for allowing the time and providing facilities for me to study for a higher degree and in particular to Dr. R. E. Reason, Mr. J. D. Barr and Mr. R. C. Spragg for their support and encouragement. Also I wish to thank Dr. P. H. Phillipson of the Department of Engineering of the University of Leicester for a number of helpful discussions of the theory of random processes.

The work embodied in this thesis is the result of the author's own research except where stated otherwise. I must, however, record my thanks to Mr. J. Jungles for his careful and painstaking work in the development of diamond styli as discussed in Chapter 3. I am also indebted to Mr. R. A. Onions for permission to reproduce the results present in Section 4.4.3 (which will form part of his Ph.D. thesis) and which allows me to display directly the application of my analysis to the problem of surface contact.

Finally I wish to thank Mrs. E. M. Whittaker for the typing of this thesis and her encouragement, and also the members of the Tracing Section of the Development Department of Rank Precision Industries Ltd. who helped in the preparation of the figures.

CONTENTS

NOMENCLATURE

A	real area of contact
a_e	elastic radius of contact
a_p	plastic radius of contact
B	ratio of wavelengths having unit to zero transmission
BR	bearing ratio
C	curvature in terms of ordinate differences
$\bar{C}$	contact
$C(\beta)$	autocorrelation function at β
$C(x)$	contact variation with distance x
c_i	ordinates of circle
E'	elastic modulus-composite
$\bar{F}$	tangential force
$F(h)$	probability distribution function at height h
$\underline{F}(\omega)$	Fourier transform
$f(h)$	probability density function at height h
f	frequency
$g(t)$	filtered output
H	hardness
$H(\omega)$	frequency characteristics of filter

$h(t)$	impulse response of filter
$\bar{h}(t)$	high-pass filter impulse response
$\underline{h}(t)$	low-pass filter impulse response
h	height value not normalised
i	count number for ordinates
j	count number for ordinates
K	ratio of ordinates to summits
K_e	envelope constant
k	height measure used in contact phenomena
$\underline{k}$	factor of asperity shape
L	length of surface
ℓ	ordinate spacing in distance
$\underline{M}$	ratio $L/2.3\beta^*$
$M(x)$	frictional force due to contacts in plastic zone
M_N	the n^{th} mean line point
m	slope of flanks in terms of ordinate differences
N	ratio of peaks to ordinates and number or ordinates
$\bar{n}$	average number of crossovers
$P(\omega)$	power spectrum
q	resolution limit of measured data

R	radius of circular part
R_a	average arithmetic departure from the reference
RMS	RMS departure from the reference
R_p	average peak height from envelope to mean line
RN_i	the i^{th} random number
RC	time constant of filter
$\bar{R}$	reaction caused by load W
$R(\omega)$	real component of $H(\omega)$
$\underline{r}$	resistivity at a peak
S	summit point
$S(\beta)$	structure function at β
$\underline{S}$	shear
S_N	sum of N distances measured in Poisson test
S	distance between apex of peaks and central ordinate
S(x)	tangential force-displacement curve
T	ordinate interval in time
t	time
t_o	co-ordinate of weighting function axis of symmetry
U(x)	Heaviside step function
U_i	distance of i^{th} peak from arbitrary position on chart
V_H	horizontal magnification
V_v	vertical magnification

W	load
$W(\beta)$	lag window
W_c	critical load
$X(\omega)$	imaginary component of $H(\omega)$
x	distance along profile chart
Y_a	uniaxial stress
Y_N	N^{th} profile point
y	vertical distance (normalised) on profile chart
y'	first derivative
y''	second derivative
y_g	gap between bodies
y_{mod}	modal value of distribution
y_s	stylus behaviour
$Z(x)$	impulse train in space
α	non-dimensional distance measurement as fraction of low cut-off value
α_i	digital filter input weighting factors
$\bar{\alpha}$	normalised equivalent of t_o
β	lag
β_i	digital filter output weighting factors
$\bar{\beta}$	normalised lag β/L
δ	impulse response
θ	local slope of surface

λ	meter cut-off value
λ_{RMS}	RMS wavelength
λ_p	Poisson density unit events
λ_s	Poisson density profile edges
$\underline{\mu}$	adhesive coefficient of friction
μ_{eff}	effective coefficient of friction
ν_p	tracking speed of pick-up
ν	Poisson's ratio
ρ	correlation coefficient
σ	standard deviation
τ	dummy time variable
ϕ	phase angle of filter characteristics
ψ	plasticity index
ω	angular frequency
$\overline{\omega^2}$	mean square angular frequency
$\underline{\omega}$	compliance

Suffix nomenclature

none	indicates profile property
superscript *	indicates peak property
superscript ^{o}	summit
superscript $\cdot$	valley
subscript $_c$	contact behaviour in general
subscript $_e$	envelope property
subscript $_g$	gap property
subscript $_s$	stylus property

1. INTRODUCTION

For many years surface finish has been recognised as an important part of manufacturing technology. The recognition of this importance has resulted in more complex methods of its measurement. Stylus tracer instruments have played a major role in this development. In very recent times an additional advantage of these instruments has been that the output, being in an electrical form, can easily be transformed into digital form for subsequent analysis. In this way, analysis of surface topography has reached a new level of sophistication requiring considerable skill in digital as well as analogue techniques. At the same time there has been only limited development of ideas about the way in which surface topography influences the functional behaviour of surfaces in engineering practice. A major role of surfaces of functional significance is their behaviour in situations involving contact and rubbing. Moreover, with only one or two exceptions, models used in the analysis of the contact and rubbing of surfaces have been very theoretical and have not been directly related to the knowledge provided by stylus instruments and the detailed analysis of their outputs which is now possible. This thesis tries to tackle this central question by relating characteristics of surfaces known to be of significance in their contact to digital analysis of the output of stylus instruments. Because the field is so large attention has been limited to properties related to surface contact and only surfaces having random characteristics have been considered. This means that surfaces prepared by mechanical methods involving the random contact of cutting elements (grinding, grit blasting, etc.) are the type of surfaces to which this work will apply.

Chapter 2 reviews, briefly, the background and literature of this chosen field of study and a more detailed published review by the author is provided in Appendix 1. Chapter 3 describes the techniques used in the later work; in particular, it outlines the methods used to analyse surface profiles (or other waveforms) when presented in digital form.

A major part of this thesis is concerned with a model in which a surface profile is presented as a random signal with a Gaussian distribution of heights and an exponential autocorrelation function. Chapter 4 provides an analysis of this model and its significance for surface contact. This theory is then compared with the results of a similar analysis of the profiles of a ground surface and the consequences of this comparison for the development of methods of characterising surfaces are discussed.

Chapter 5 provides a theoretical analysis of the movement of a second body over the surface of a body having a random profile, without any deformation of either. The more complex question of the consequences of rubbing surfaces under load has been studied experimentally and the results of this work are reported in Chapter 6; the extent to which these observations can be treated by the same form of analysis is discussed. In Chapter 7 the generation of random surfaces by mechanical methods is considered; the major object is to explore (in an elementary, but fundamental manner) the extent to which the observed characteristics of profiles of random surfaces might be expected on theoretical grounds. Finally Chapter 8 draws some broad conclusions from the work described in the thesis and outlines the directions in which the subject may develop.

2. SURFACE TOPOGRAPHY AND SURFACE CONTACT

2.1 Introduction

Surfaces are becoming more and more important. The requirements of modern technology are placing an ever-increasing burden upon the surface and the surface layers of components. This calls for a greater understanding of the nature of surfaces, of their measurement and classification, of their features and of their control in manufacture. In order to be able to decide on these properties, both physical and topographic, which are most likely to be of use in controlling manufacture and predicting functional behaviour it is necessary to examine in detail some of the various functions to which surfaces are required to perform.

A comprehensive review of the subject covering the most important aspects of this problem from function through to geometrical classification is given in Appendix 1. This chapter will select, for brief discussion, those aspects of particular significance for the work described in this thesis.

2.2 Surface contact

In order to be able to understand the behaviour of surfaces in friction and wear a consideration of the mechanism of solid contact is essential. The nature of the contact either dry or through lubricant films, together with the physical properties of the materials, will determine, to a large extent, the performance of the surface; for example its suceptibility to damage in sliding contact or its ability to run-in.

Theories of surface contact are derived mainly from the equations for a single contact region usually represented by the contact between a smooth sphere and a flat surface. At light loads there is elastic deformation of the two bodies. If the radius of the sphere is R, the load W, E_1, E_2 and ν_1, ν_2 are the values of Young's modulus and Poisson's ratio for the two materials respectively then using the Hertz equations the radius a_e of the area of contact is given by

$$a_e = \left[\frac{3}{4} WR \left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \right) \right]^{1/2} \qquad (2\text{-}1)$$

which reduces to

$$a_e = 1.11 \left(\frac{WR}{E}\right)^{1/3} \qquad (2\text{-}2)$$

when $E_1 = E_2$ and $\nu_1 = \nu_2 = 0.3$

At very heavy loads the size of the contact region is dominated by the plastic behaviour of the material and for circular contact area the radius of the contact area is given by a_p; where

$$\pi a_p^2 H = W$$

or $$a_p = \left(\frac{W}{\pi H}\right)^{1/2} \qquad (2\text{-}3)$$

where H is the flow pressure or hardness of the softer of the two materials.

The load at which plastic flow first occurs can be derived using the equations for an elastic contact. At this load the maximum shear stress just reaches a value of $Y_a/2$ where Y_a is the yield stress in uni-axial mode. Then using the relation

$$H = 2.7Y_a \qquad (2\text{-}4)$$

the load at which the onset of plastic flow occurs can be calculated.

A measure of the load at which the transition from elastic to plastic deformation occurs is that value such that $a_p = a_e$. Equating these from equations (2-2) and (2-3) gives

$$W_c = 55 \frac{R^2H^3}{E^2} \qquad (2\text{-}5)$$

Then, to a reasonable approximation, the following conditions apply. Fully elastic conditions occur at loads < $W_c/15$ and the first plastic flow occurs at this load. Fully plastic conditions occur at loads > $40 \frac{W_c}{3}$. In between there is a transition region between the onset of plastic flow and complete plastic flow.

Theories of surface contact are concerned with the behaviour of individual contacts and with the subdivision of the total real area of contact into multiple contacts which occurs when rough surfaces are used. Much of the earlier work on surface contact Holm (1958), Bowden and Tabor (1954) and Merchant (1940) was aimed at producing a rational explanation of Amontons' Laws of Friction (Amontons 1699) which are that the frictional force is

(a) proportional to the load, and (b) independent of the area of contact. The first important point realised was that the real area of contact is much smaller than the apparent area calculated from the dimensions of the parts. This together with the formulation of the adhesion theory of friction enabled the laws to be explained. The important assumption was made that the real area of contact arose from plastic deformation of asperities under the load W. The area of contact A would then be equal to $\frac{W}{H}$ where H is the hardness of the softer material. The adhesion theory of friction also assumes that the frictional force arises from the force required to shear the junctions at this real area of contact. Hence the frictional force is proportional to real area of contact, which in turn is proportional to the load. This theory did not attach much importance to the surface finish because it plays no part in determining the severity of the contact conditions, but Archard (1957) pointed out that although plastic flow could be expected to occur on the first few passes of two contacting parts in relative motion it would not continue indefinitely, some equilibrium state would occur when the asperities could support the load elastically. He then went on to show that Amontons' Laws could be explained using elastic deformation theory providing the average contact size remaind constant with load. This was a direct result of having an increase in the number of contacts with load, a point which required more than one scale of size of asperity on the surface.

In order to decide which of these two deformation modes, elastic or plastic, occurs in practice, it is necessary to consider the nature of the surface geometry.

2.3 Surface topography and its measurement

In this section we will be concerned only with the geometrical properties of the surface - some other properties which are significant in the functional behaviour are discussed in Appendix 1.

The physical size of the marks left on the part by the manufacturing process is very small. Four orders of magnitude in the size of the roughness values exist, ranging from 0.1 μm for polishing, to almost 1 mm for shaping; a typical size would be about 5 μm for turning. In the horizontal spacings the range is about three orders of magnitude from a few millimetres down to a micrometre or thereabouts; as above, a typical value is about 20 μm for a turned surface. These small sizes and wide range of values make assessment of the surface geometry difficult by eye or thumbnail consequently a wide variety of instrumental methods have been devised to more accurately assess the surface geometry. These range from optical, pneumatic and capacitative techniques, for example, to the commonly used stylus tracer instruments.

The optical methods usually used are either based on a microscope (Martin 1967) or an interferometer (Tolansky 1970) or both. Normal viewing under a microscope gives information over an area; the only information in the vertical plane is obtained by use of the focusing mechanism or by the use of oblique lighting which causes shadows on the surface from which estimates of height can be made.

Vertical information of high accuracy can be obtained at the expense of some loss of area information by the use of interferometry; this usually involves the positioning of a reference plate on top of the surface either parallel to, or at an angle to, the general direction of the surface. Contour lines of the surface are produced by this means. Another technique due to Linnik enables a greater numerical aperture of the fringe viewing optics and hence better resolution to be obtained by positioning the reference flat in a remote place away from the surface. (Reason 1969). Better results can also be obtained by use of multiple beam interference (Tolansky 1970) in which both the surface and master plate are coated. Other techniques such as Nomarski interference contrast and phase contrast can be used to advantage on smooth surfaces. In recent times the assessment of the surface geometry using goniophotometric techniques has been gaining popularity. (See Bennett and Porteus 1961, Davies 1954 and Reneau and Collinson 1965).

However, although these latter methods can be made to give satisfactory results for some surfaces they have, as yet, been impossible to make universally useful for the assessment of the surface texture. The other more conventional optical techniques, although useful, have not found extensive use in workshops because of the difficulty in getting, cheaply, a number as the output which relates directly to the surface finish.

Pneumatic and capacitative techniques have been used over a number of years but they have found little general use. In pneumatics, for instance, the sensitivity is inadequate for many uses whereas in capacitative techniques, where the capacitance between a reference plate and the surface is measured in order to assess the surface roughness, one of the difficulties is found in measuring any surface other than those having a flat shape; a different shaped reference capacitor plate being required for each shape of surface.

It is, however, stylus instruments which have become the most widely used for the assessment of surface texture. This is because of their ease of use, and convenient and unambiguous output. (See Reason et al 1944 and Reason 1956)). In fact, in recent years, the use of the stylus instrument as a research as well as an inspection tool has been increasing mainly because of the advent of digital techniques (Reason 1964(a). Consequently this thesis will be concerned only with the results obtained from stylus instruments and in particular those results obtained using digital techniques.

One of the features that has emerged in the investigation of surface geometry using stylus instruments and digital methods has been the importance of analysing good quality data in the computer. Before any useful comparison of results can take place the quality of the data used must be assured. The author has been very aware of this point and has taken steps to ensure the quality of the input data to the computer. This has taken two forms. First, a method has been devised in which the shape of the stylus itself can be accurately assessed (Jungles and Whitehouse 1970). Second, is the processing of the digital data prior to analysis in the computer. In particular, a digital filter has been devised which has optimum attenuation and phase characteristics and which allows the removal, from the digital data, of the extraneous long wavelength components often met with in surface profiles (Whitehouse 1968). These techniques are dealt with in Chapter 3.

One of the main interests in surface finish research at the present time is that of trying to develop a typology of the surface geometry so that a completely adequate classification of the surface can be made without resorting to the existing technique of specifying the R_a value together with a statement of the manufacturing process. This existing technique, although very useful, is becoming increasingly unsatisfactory in some respects because (a) it cannot adequately predict the behaviour of the surface in some of the more stringent modern engineering functions, and (b) it can prove restricting to the production engineer who wants complete freedom in the choice of manufacturing process. Before discussing some of the work that has been done on topographic typology it is proposed first to consider some of the fundamentals which must be involved; because any discussion of typology naturally leads to statistics we will examine initially some of the statistical considerations.

Surfaces can be random or deterministic or, more usually, a mixture of both. For a complete specification of a general random process, high order joint probability density functions are needed (Bendat 1958). However, in practice a second order joint probability density function will suffice. From this both the ordinate height distribution and the autocorrelation function, and hence the power spectrum, can be found. Because the second order probability density function is, in general, not known, the autocorrelation function and the ordinate height distribution can be conveniently used to define the statistics of the profile. In the many practical instances where the statistics of the process is Gaussian, or thereabouts, then the normalised autocorrelation function and the RMS (or average value R_a) completely define the profile. One of the main reasons why these are a good basis for considerations of typology is that the autocorrelation function has the useful property of being able to separate the random from the periodic components of a waveform. Such has been the usefulness of these statistical parameters that they have been used in many different fields.

One practical point about any typology is that the parameters used should, from an instrumental point of view, be kept simple and cheap. Furthermore for a parameter to be useful in typology it must be at one and the same time both discriminatory to distinguish between one surface and another while being reliable enough not to produce wildly varying values over the same surface. Another point which tends to be neglected is that a typology should be capable of taking into account not only the overall statistics of the surface but also statistically unpredictable freak events. This is one region where correlation techniques are of little use. Such behaviour is difficult to predict even using statistics (Gumbel 1959).

Turning back to the statistical aspects most of the present-day parameters of surface geometry are basically estimates either of the height distribution or of the autocorrelation or mixtures of the two. The importance of the ordinate height probability density function in surface metrology was first realised by Abbott and Firestone (1933) who proposed the use of a curve showing how the ratio of metal to air changed with the height of a hypothetical flat plate lapping away the surface from the highest peak to the lowest valley. This curve is generally referred to as the bearing area (or ratio) curve; it is in fact one minus the ordinate height distribution function. Pesante (1963) proposed a classification according to the shape of the ordinate height density function. He found it more useful than the bearing area curve because it was more discriminating. Reason (1964(b)) proposed the use of the consolidated bearing area curve together with the high spot count to classify the surface, and Ehrenreich (1959) suggested that measurement of the slope of the bearing area curve could be useful.

The R_a and RMS values are essentially estimates of the scale of size of the ordinate height distribution - as indeed, also, are any peak height measure such as the maximum peak-to-valley height. Some attempt, however, usually has to be made in peak height measures to preclude freak events. To this end the Swedish Standard considers the difference in height between the 5% and 90% bearing area percentages and the British Standard considers the difference in height between the five highest peaks and five lowest valleys. Other features of the density function can also be considered to be useful, especially in demonstrating wear, for instance, the skew. Al-Salihi (1967) in fact proposes that in addition to the RMS value the third, fourth and higher central moments should be considered. Unfortunately these are difficult to measure reliably.

The fundamental reason why the height distribution itself is of limited value is that it contains no information about the bandwidth of the profile waveform. Before considering methods that have been evolved in answering this problem directly we will first consider those methods of classification which involve the derivatives of the surface profile. Myers (1962) recommended the use of the RMS values not just of the profile itself, but the RMS values of the profile slope and second derivative, together with a directional parameter. Other investigators have proposed the use of either one or more of the derivative parameters. Peklenik (1963) considered the value of the standard deviation of the slope as a convenient estimate of the autocorrelation function. The use of the distribution of the slope has been reported by others including Kubo (1965), Nara (1962). Nara suggested that the R_a value and the mean slope value could be used for specifying a surface on a two-dimensional graph. He maintained that by doing this he could estimate both the drop-off of the autocorrelation function and the high spot density.

One of the important practical points concerning the use of these highly discriminating parameters like slope or curvature measurement is that they tend, by their very nature, to reduce the effective signal to noise ratio, i.e. extraneous short wavelength noise tends to get amplified. One way out of this problem is to introduce a short wavelength filter. This will be mentioned later on concerning the work of Spragg and Whitehouse (1971).

Many people have pointed out that there is a functional need for a spacing type of parameter, for instance in the sheet steel industry (Butler and Pope 1968). Some examples of parameters that have been used are the number of crossings at a given height (Reason, 1964(b), Pesante 1963, Peklenik 1963). The number of peaks in a given length has also been used. Sometimes, for example, in the American sheet steel industry, the definition of a peak is different from the normal one; they insist on the valley following the peak being more than a fixed distance below. A more recent measure is the average wavelength introduced by Spragg and Whitehouse (1971) which takes into account the size of all the harmonics as well as the dominant spacing.

Any classification must be a condensation of information. Of all the thousands of bits of information contained in a typical waveform only a few are going to be needed for any given function. (Ultimately we require one piece of information - the answer to the question "Will it perform satisfactorily?" - but this begs the question of where this information is to be found). This is where the autocorrelation function is useful because it represents a useful condensation of the information in the waveform.

Wormersley and Hopkins (1945) were the first to put forward effectively the autocorrelation function (in a time series form) as a useful measure of surface texture followed by Linnik (1954) and Nakamura (1960). However, it was Peklenik (1967) who proposed the further condensation of the autocorrelation into groups suitable for use as a classification system. He proposed classifying the autocorrelation function to decide into which group it best fitted. The surface was then typified by the number of the group. Thus he was able to present surfaces made by different processes on a typographic scale which comprises:

Group 1 - Cosine or steady valued.

Group 2 - Exponential decay plus cosine.

Group 3 - Exponential decay modulating a cosine.

Group 4 - Complex combination of groups 2 and 3.

Group 5 - Exponential decay.

In this classification, Group 5, for instance, (first order random surface) is typical of grinding honing etc., whereas Group 3 (second order random surface) together with Group 2 are more typical of single-point cutting processes like shaping, turning etc. Group 1 is purely deterministic and does not occur on practical surfaces.

As a further subdivision of each group, Peklenik (1967) introduces the correlation length and the correlation period; the former measuring the rate of decay of the autocorrelation function and the latter measuring the spacing of the dominant periodicity. Although this classification system is a major step forward in the specification of surface texture it has certain difficulties in its application. These, and some proposed amendments to include the classification of the ordinate height distribution are discussed in Appendix 1 (Whitehouse 1970). The extension of surface assessment to three dimensions by Peklenik and Kubo (1968), McAdams et al (1968), and others is also discussed in Appendix 1.

2.4 Surface topography and surface contact

From what has been said in Sections 2.2 and 2.3 it is clear that a great deal of time and effort by researchers has been put into the fields of surface contact and surface topography. In the field of surface topography, progress has been particularly rapid over the past few years. However, these two fields have developed practically in isolation. The result is that many questions are still left unsolved; in particular the influence of surface finish upon behaviour involving contact, such as wear and friction, still remains largely unknown. An important example is in determining the mode of the deformation that occurs under different conditions when two bodies are contacted. It used to be thought (Bowden and Tabor 1954) that the asperities were always plastically deformed upon compression. More recently it has been recognised that surface contact must often involve an appreciable proportion of asperity contacts under which the deformation is partially or completely elastic (Archard 1957); consequently, the surface finish must play a large part in determining the proportion of elastic and plastic deformation. The great need at present, therefore, is the bringing together of the theories of contact with the characteristics of surfaces as determined by surface metrology.

One of the few, and perhaps most successful, attempts to bring these disciplines together has been by Greenwood and Williamson (1966). They assume that the surface is made up of a Gaussian distribution of asperities of standard deviation σ^* and that upon contact, say with a flat plate, only the upper tips of the asperities actually make contact. They assume that all asperities have a radius of curvature R at the tip. They assessed the probability of plastic deformation of the asperities using this model. In fact there is always a finite chance of plastic flow using this model; however, one important conclusion that Greenwood and Williamson came to is that the probability of plastic flow depended very little on the actual load but is critically dependent upon a plasticity index ψ given by

$$\psi = \frac{E'}{H} \left(\frac{\sigma^*}{R}\right)^{1/2} \qquad (2\text{-}6)$$

where E' is the composite Young's Modulus and H is the hardness. Unfortunately although this equation represents a considerable advance, their theory has its limitations. They do not take account of the existence upon surfaces of superposed asperities of differing scales of size. Also the plasticity index assumes that the deformation of each of the asperities is independent, consequently the plasticity index has significance only if it is applied to the main long wavelength structure of the surface. Also the theory does not take account of the distribution of peak curvatures always found on surfaces. One final important point is that their theory only relates contact phenomena to peak characteristics and not the characteristics of the profile waveform used in practice to assess the surface texture. The need, therefore, is for further steps to complete the bridging of the gap between contact theory and modern methods of measuring and classifying surface texture.

3. TECHNIQUES

3.1 Introduction

A large part of this thesis is concerned with information derived from Talysurf profilometer instruments presented in digital form. The significance of this technique in the development of the subject has been already discussed in Chapter 2 and Appendix 1. Broadly speaking its advantages lie in the range of processing operations which become available when data are available in digital form; these operations are possible because of the availability of fast digital computers. For these same reasons similar techniques have been used in the present work for the analysis of experimental data presented in Chapter 6. The experimental data include values of the frictional force and the displacement of one body when it moves over another (which we shall describe as the "ride"), as well as the surface profiles of the rubbing bodies. These measurements, and the details of the apparatus used, will be described at the appropriate point in Chapter 6. However, at this stage it is relevant to explain that, for obvious reasons of convenience, these measurements have been made using a Talysurf stylus and its associated circuits as a displacement transducer. In this way the whole range of experimental data presented in this thesis (and not merely the surface profiles) has been presented in the same digital form for subsequent analysis.

To allow the description of the later work to proceed without interruption, this chapter contains a description of the techniques used to transform the instrument output into digital form suitable

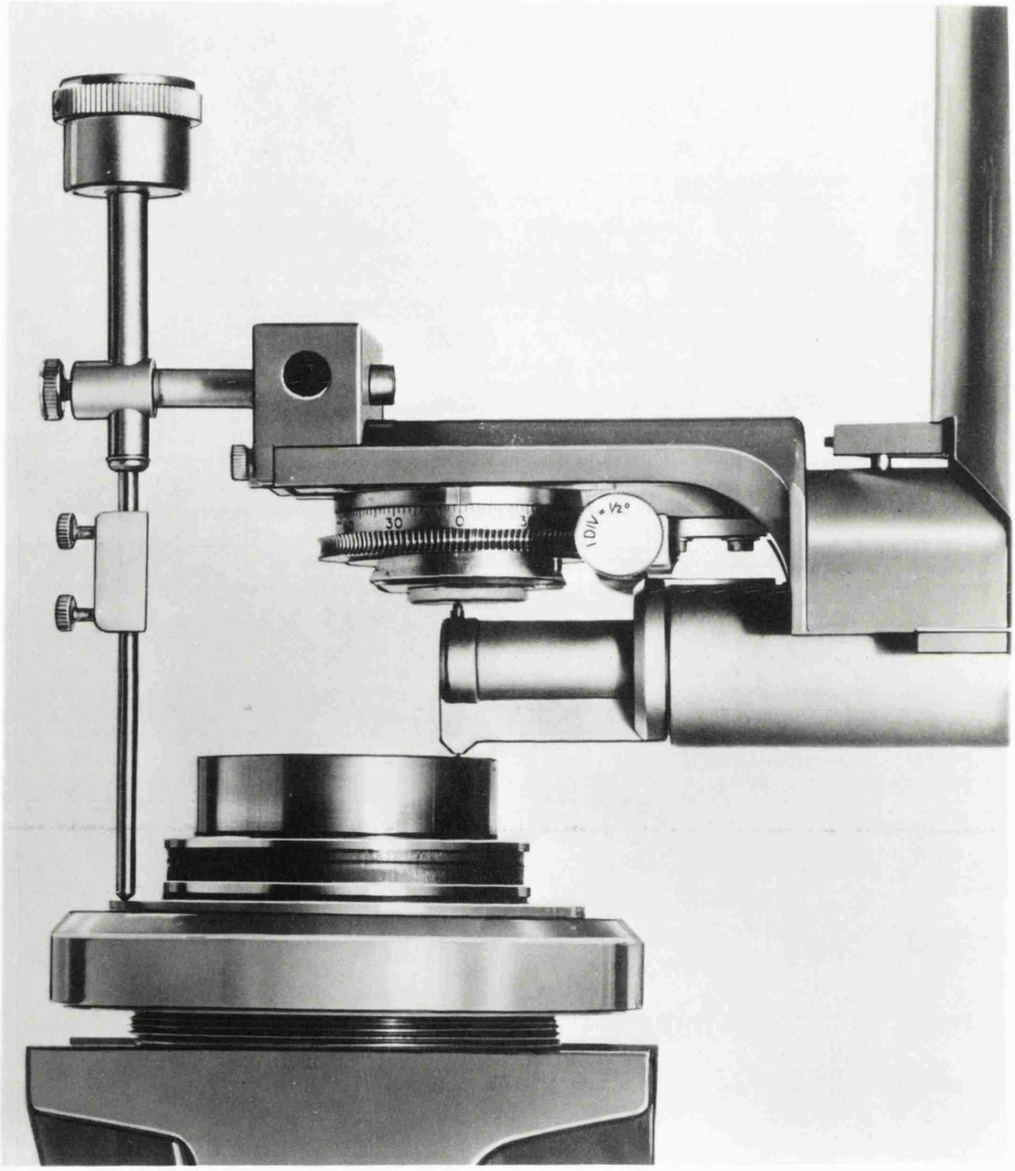


Figure 3-1. Talysurf Pick-up showing Datum Attachment.

for use with the computer. The computer techniques used in the analysis of this data are then discussed. Because of the number of different programs used in this work this discussion of computer analysis has been, necessarily, confined to the broad principles involved. However, as an example, one program has been selected for more detailed explanation. Most of these techniques of analysis are also applicable to the results obtained in Chapter 7; here surface profiles, in digital form, are obtained from simulation of the mechanism of generation instead of from the output of the Talysurf instrument.

To set this work in its proper context, this chapter opens with a brief discussion of the Talysurf instrument and its successor the Talystep, which has the potential for development as a high resolution profilometer. The question of instrument resolution has played an important role in the work described in Chapter 4 in which divergence between the theoretical model and the experimental measurements may be in part attributable to stylus resolution. The stylus shape plays an important role in instrument resolution. For this reason, the author has been involved in some detailed examination of the shape of styli used in profilometry and their method of manufacture. A brief account of this work is included in this chapter.

3.2 Stylus Instruments

The basic instrument used in this work has been the Talysurf 4 which is a stylus tracer instrument used extensively in industry for the measurement of surface texture. In the operation of this instrument a sharp stylus is tracked slowly across the machined surface. The up and down movements of the stylus are amplified and registered on a meter and a recorder. The stylus is usually a diamond pyramid typically of tip dimension 2.5 μm in the direction of traverse. The pick-up element upon which the tip is fixed is constrained to have only one degree of freedom and as a result of this the up and down movements which are communicated to it by the stylus represent an accurate geometrical representation of the surface itself in the one track over which the stylus is passing. (Detailed descriptions of the electronics are contained in the handbook). It must suffice here to say that the instrument transducer is of the inductive type. In this an armature which is connected to the pick-up element moves within a coil according to the movement of the stylus. The coil itself is part of a bridge circuit being fed from a 10 kHz carrier signal. The changes in the position of the armature cause different inductance in the two halves of the coil which has the effect of converting the stylus, and hence armature movement, into an amplitude modulated voltage which can then be processed by filters and other appropriate circuits. An essential feature of the mechanical principle is that the surface roughness is measured relative to a straight smooth optical flat which is supported above the stylus, figure 3-1. It is the movement of the stylus <u>relative</u> to this optical flat which constitutes the measured surface roughness. For convenience a crude datum called a skid, or alternatively a shoe, can sometimes be used. These take the form

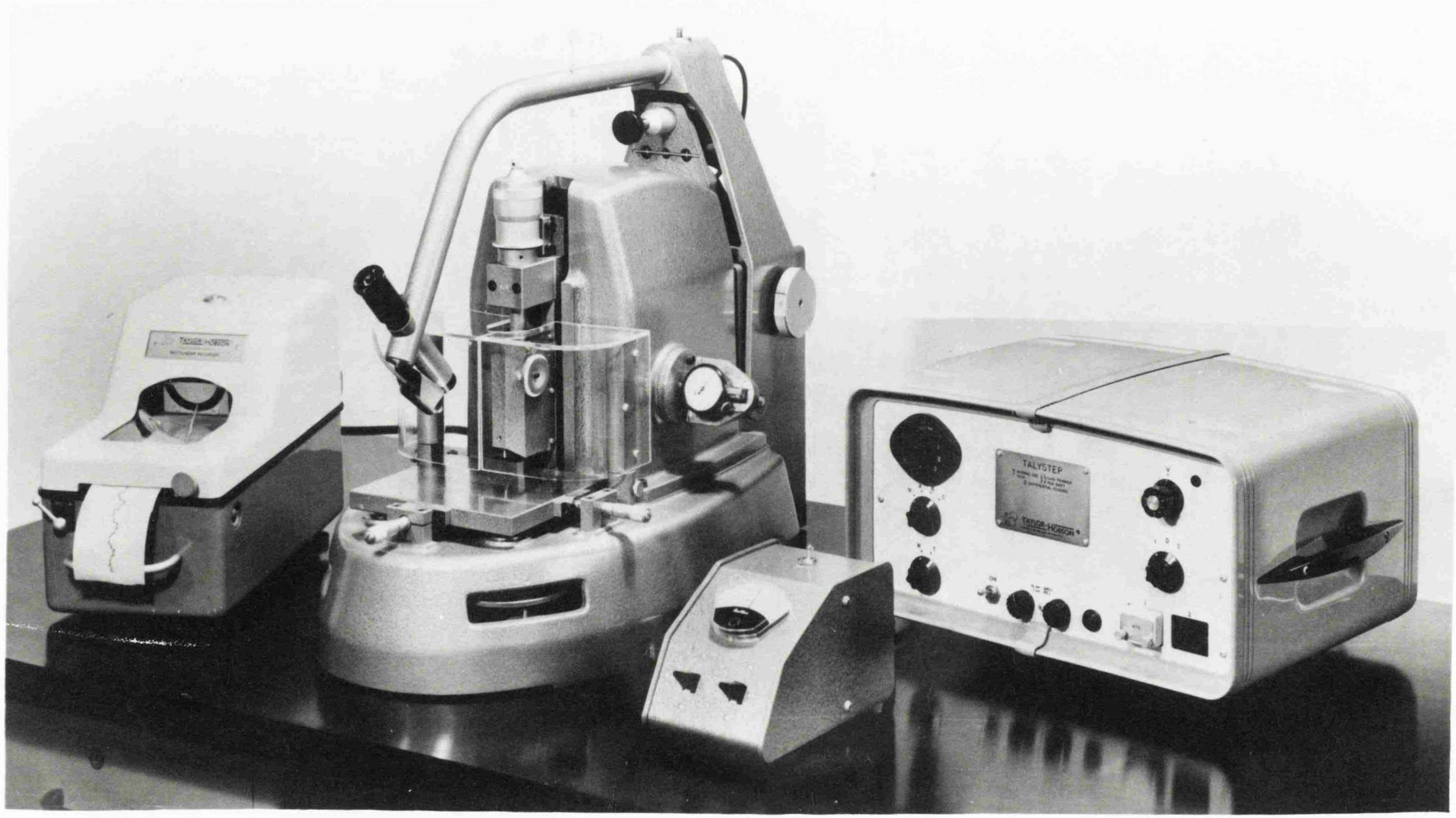


Figure 3-2. The Talystep showing the instrument Amplifier and Recorder.

of a blunt foot which rests on the surface and to which the body of the pick-up is attached. The mechanical reference relative to which the stylus movement is measured is then taken as the difference between the skid vertical position and the stylus. Upon being moved, because the stylus is so very much sharper than the skid, an approximate profile waveform is generated. This technique was not used during these experiments.

The Talystep, figure 3-2, is another type of stylus tracer instrument. It works on a principle which is basically the same as that of the Talysurf. However, there are important differences, mainly mechanical. First and foremost is that the stylus load, instead of being 100 mg as in Talysurf, can be varied and reduced down to loadings as low as 0.5 mg - a feature which makes possible the use of very sharp styli. Also the maximum magnifications obtainable are different from that of a Talysurf. Instead of a 10^5 top vertical magnification, for the Talystep it is 10^6; horizontally it is 2000 instead of 500. The Talystep, however, has a much shorter traverse length.

The two greatest advantages of these tracer instruments are (a) that the output is in the form of an electrical signal which can easily be processed, and (b) that they are very convenient and easy to use.

3.3 Measurement of stylus shape

3.3.1 The significance of stylus shape

For most practical applications of tracer type instruments the fact that the finite size of the tip of the stylus must filter out some of the short wavelengths on the surface is not important. Neither is it important that the minute structure and shape of the tip itself are not precisely known. However, in some cases where ultra-fine surface finish is being measured or detailed investigation of the fine structure on surfaces is being carried out then it becomes not only essential to use a much sharper stylus, it also becomes very important that the geometry of the tip is known. To use a very sharp tip of the order of 0.1 μm dimension is not practical for two reasons. First, the loading of the stylus taken with the smaller tip would take the pressures at the tip well behind the yield point of most metals. Second, the relatively coarse movement of the Talysurf would not be conducive to the maintenance of the fragile tip. However, use of the Talystep, is practical and, for the purpose of the investigation described in Chapter 4, is an admirably suitable instrument. Use of such small styli not only gives instrumental problems it also highlights the problem of the measurement of the tip geometry. Obviously the first technique that comes to mind is that using optics.

3.3.2 Optical Microscopy

For many years the nominal tip dimension used in conventional stylus tracer instruments have been of the order of 2.5 μm for

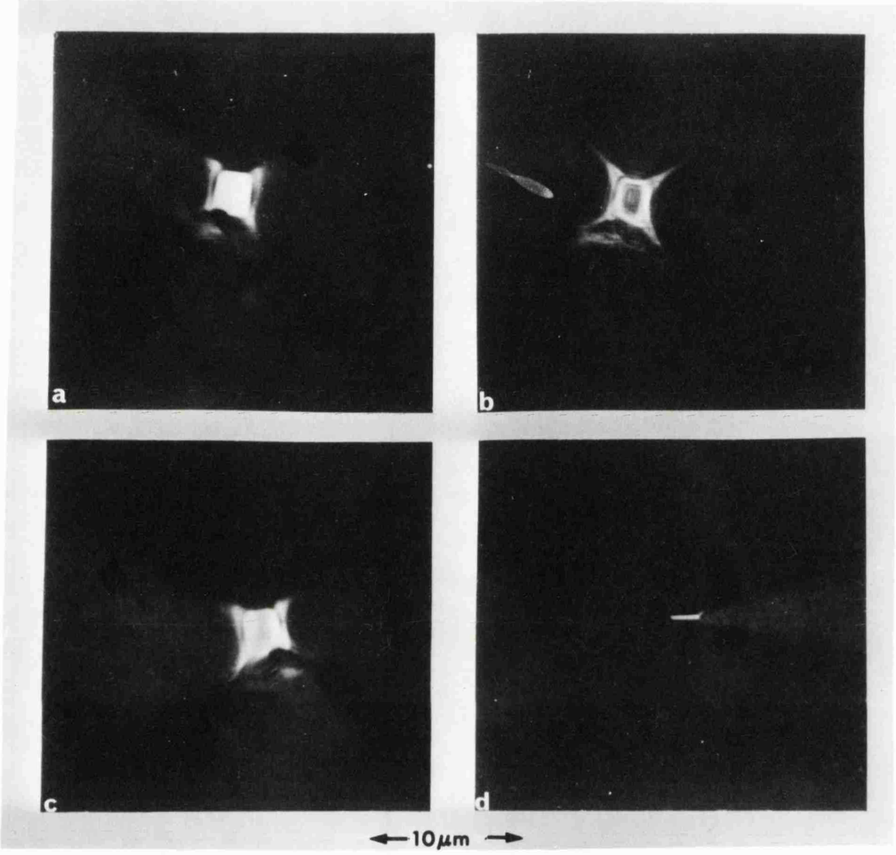


Figure 3-3. Optical methods of stylus measurement.

(a) 2.5 μm stylus using normal bright field.

(b) Stylus as in (a) but using phase contrast.

(c) Stylus as in (a) but using Nomarski Interference.

(d) Sharp chisel stylus using normal bright field.

normal surface profilometry applications. Dimensions of this size can, to some extent, be examined optically. For instance in figure 3-3 a diamond tip having the dimension of about 2.5 μm square is shown as viewed with the different optical techniques of (a) normal bright field, (b) phase contrast, and (c) Nomarski interference. It will be seen that it is difficult to determine the real dimension of the tip itself let alone any fine structure that may be present at the tip.

For the work described in Chapter 4 it therefore becomes necessary to devise a method of measuring the microgeometry of the stylus tip. Another problem emerged as a result of this exercise in Chapter 4; this was the manufacture of diamond styli of extremely small dimension, sufficient for the requirements of the investigation. In the solution of both of these problems the author gratefully acknowledges the significant contribution of Mr. John Jungles of Rank Precision Industries Ltd., Metrology Research Laboratory. These details are covered in Sections 3.3.3 to 3.3.5. A published paper on these techniques is included in Appendix 3 (Jungles and Whitehouse 1970).

3.3.3 Scanning Electron Microscopy

From what has been said concerning the various optical techniques and judging from the pictures shown in figure 3-3, it would seem natural to apply the scanning electron microscope technique to the measurement of the stylus tips. This is because of its obvious advantages over light microscopy. These advantages are (a) the depth of focus is increased because of the long focus

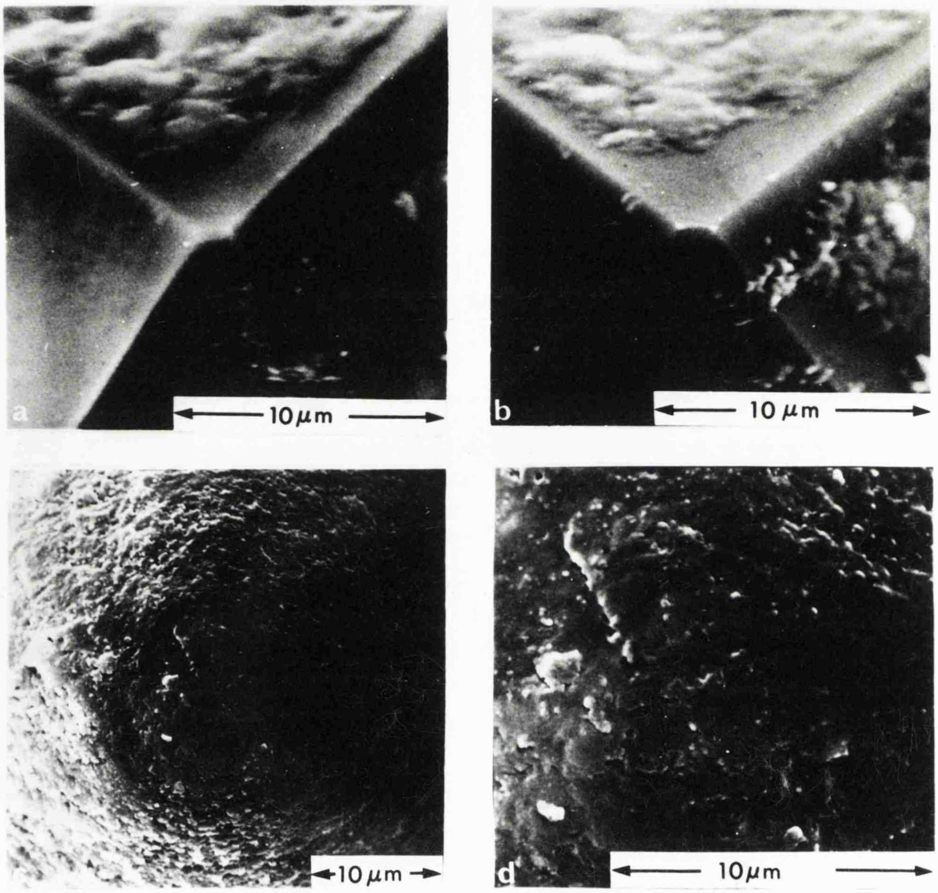


Figure 3-4. Scanning electron micrographs of styli.

(a) Sharp chisel stylus.

(b) Stylus as (a) but aluminium deposit.

(c) Gramophone stylus.

(d) Stylus as in (c) but at a higher magnification.

magnetic lens which keeps the beam divergence small, and (b) the increase in the useful magnification which is made possible by the very small equivalent wavelength of the electron beam.

Unfortunately it was found that the measurement of diamond tips by secondary emission techniques was not completely satisfactory. An example is shown in figure 3-4(a) which is a scanning electron micrograph of the chisel shaped stylus shown in figure 3-5(b). It can be seen that although there is a considerable improvement it still leaves a lot to be desired. This is due to two things, the first being that the diamond is an insulator and the second is the extreme nature of the geometrical shape of the tip. Secondary emission from an insulator such as diamond can be a complex phenomenon because of the absence of free electrons. An insulator will not lose energy as in a metal by interaction with free electrons in the conduction band. Primary electrons will only lose energy by interaction with the valence electrons and, unless the insulating object is a thin film on an electrically conducting base, a space charge forms in the material. It is possible for an insulator to emit more electrons than were introduced giving rise to a net change in charge which causes charge to migrate towards peaks or other sharp boundaries.

Attempts to reduce this effect by deposition of a thin deposit of conductive materials like aluminium, gold or carbon have not much effect as seen in figure 3-4 (b).

Consequently it became clear that although scanning electron techniques are suitable - even for diamonds having little extreme geometrical shape, as in the gramophone styli, figure 3- 4(c) and (d) it is not really satisfactory for the diamond tips in question.

3.3.4 Transmission electron microscopy

In this technique the electron beam is passed directly through a replica of the object to be measured. In the case of diamond tips this in itself presents problems because any replica must not only show the details of the tip clearly but must show enough of the overall geometry of the stylus to enable it to be found when in the microscope. Usually all sorts of various shapes and markings litter the view. Some sort of positive identification is essential. For this reason the natural single-stage replication materials like gelatine, collodion and formvar and some two-stage techniques which utilise wax, gelatine, acetates and metals as the first stage failed.

Glass as a first stage replica material was tried. Glasses differ in properties from the organic polymers which are usually used for replicas in electron microscopy. Whereas the polymers are partly crystalline and only to some degree amorphous, glass is completely amorphous which means that any replica of the tip is likely to be faithful - at least for a limited time. One other advantage of glass is the ease with which a second replica of carbon can be released from the glass. By these techniques the required results were achieved.

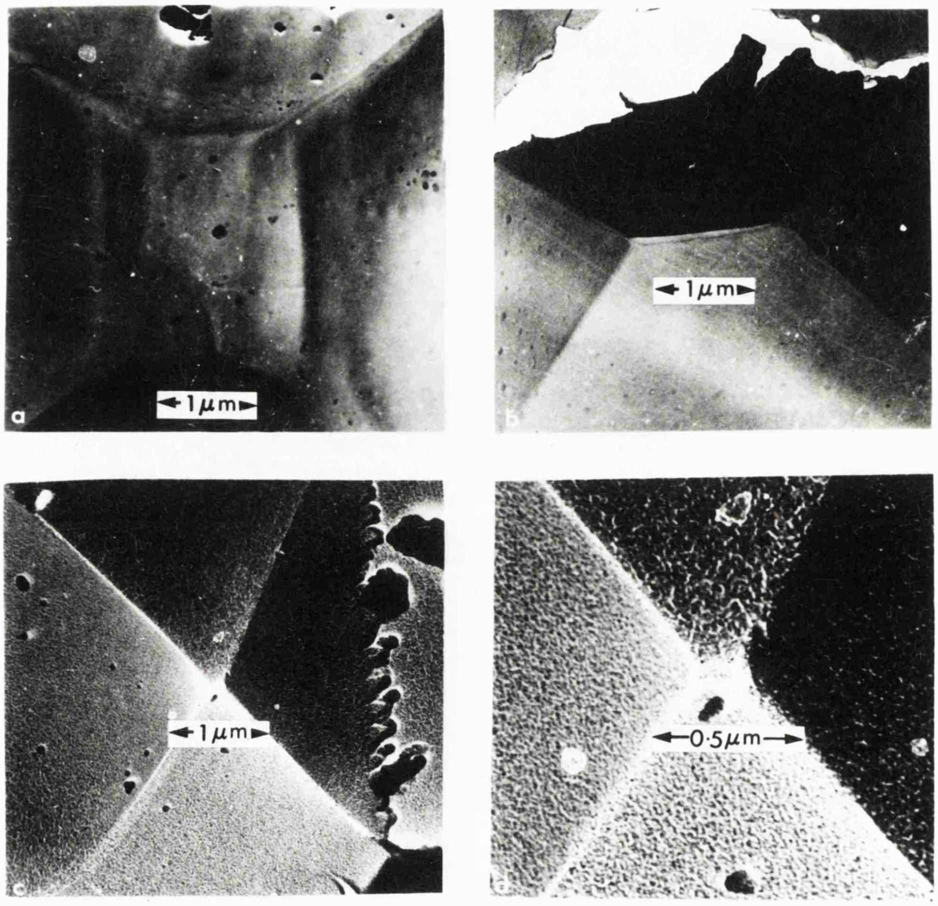


Figure 3-5. Transmission electron micrographs of styli.

(a) Stylus shown in Fig. 3-3(a).

(b) Stylus shown in Fig. 3-3(d).

(c) Ultra-sharp stylus.

(d) Stylus as in (c) at a higher magnification.

No damage to the tip could be expected through indenting the glass because the relative hardness of diamond and glass is about the same as for tungsten carbide and zinc, so the tip is most unlikely to be damaged. One result which supports this statement is that no changes were observed in a succession of replicas taken from the same tip.

Practically the procedure was as follows:

A small rig was constructed for use on a standard instrument (Talystep) which enabled a known load to be applied while the diamond was resting on the glass. A given small force of about 1 grm was then applied smoothly for about a second. This process was repeated many times within a small region. Having such control enables many such indentations to be applied in the same way.

Carbon was then deposited onto the indentation from one or two directions at right angles up to a thickness of about 0.03 μm to give rigidity. This was subsequently shadowed with gold or platinum to a depth of about 5 nm. The replica was then floated from the glass by immersion in water. The carbon replica had then to be removed and placed on a standard electron microscope grid. The depth of carbon used was necessary to give rigidity. Rigidity of the replica was sometimes difficult to maintain as is shown in figure 3-5(b) which shows a typical replica fracture.

Typical results are shown in figure 3-5. Picture (a) is that of the tip shown in figure 3-3, (a), (b) and (c) showing the clear detail. Picture (b) is that of the stylus shown in

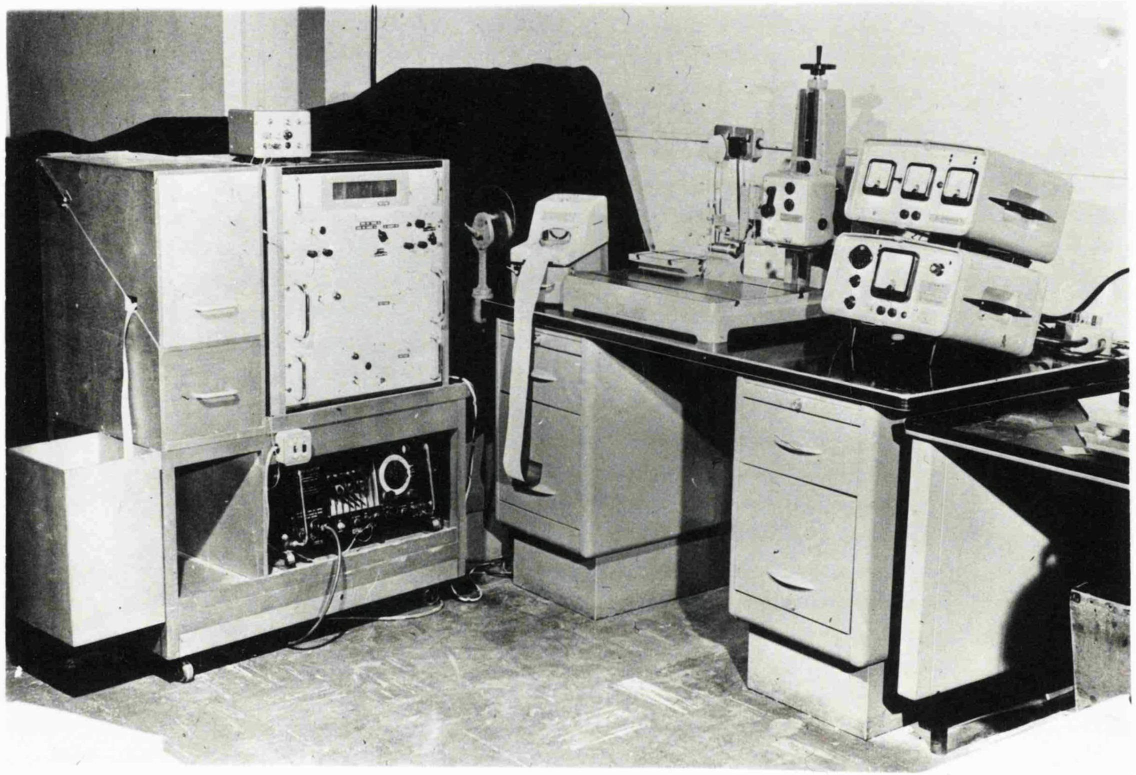


Figure 3-6. Data Logger coupled to Talysurf 4.

figure 3-3(d), and finally figure 3-5 (c) and (d) are micrographs of an ultra sharp stylus.

3.3.5 Manufacture of a stylus for ultra high resolution

A previously mentioned a sharp stylus was required for testing the limits of the theory in Chapter 4. The stylus that was made is shown in figure 3-5(c) and (d). It was made by lightly loading an ordinary stylus against a slowly rotating cast iron disc charged with one micron diamond paste. Arrangements were made so that the diamond could be turned accurately through 90° periodically. This method differs from the conventional techniques in that it uses much smaller loads and much slower speeds. Although this results in long periods of polishing, of the order of days, the final result justified the delay. Using this technique it was also possible to make styli of sharper angle than the conventional 90° pyramid but the combination of sharp tip and acute angle make it mechanically fragile to use.

3.4 Digital techniques

3.4.1 Analogue/digital conversion

In addition to the basic analogue instrument, the Talysurf or Talystep, a data logging system has been used (figure 3-6). This comprises a Solartron A/D converter and serialiser which intercept the Talysurf or Talystep signal immediately after the recorder amplifier. The digital signal is then fed to either a

Data Dynamics 110, 5-channel paper tape punch, or alternatively to a Facit 8-channel paper tape punch. Both of these systems are capable of about twenty digital measurements per second. In what follows the digital value of a sample of the height of the waveform will be referred to as a profile ordinate or simply an ordinate. Each ordinate consists of three decimal digits together with a fixed character symbolising the end of the ordinate (or word).

The Talysurf signal was arranged so that the total width of the recorder paper corresponded to 999 units on the paper tape; zero being on one edge of the recorder paper rather than in the centre. This saves a polarity character.

Every ordinate was represented, therefore, by a number between 0 and 999 which saved the use of a character for a decimal point. This meant that each measurement of the waveform had a resolution of 10 bits; the relative accuracy of successive ordinates was therefore 10 binary bits. (This does not mean that the waveform itself was accurate to this value - a figure of about 2% would be realistic.

Digital techniques were preferred over analogue because they are (a) intrinsically more accurate, (b) more versatile, (c) better for storage and display, and (d) quicker and cheaper in the long run. These will be explained briefly:

(a) It is difficult to get analogue instruments that operate on the Talysurf electrical waveform to better than one percent, as just stated 10 bit accuracy is easily possible digitally. Other points on this topic will be referred to later.

(b) By merely writing a program any parameter of the waveform can be measured. This is usually not too difficult. It is usually much more difficult to do the same things by analogue techniques.

(c) The form in which the data is obtained in the digital technique makes it more suitable for storage and retrieval than the usual analogue techniques. In the system developed here the paper tape output was transcribed onto magnetic tape. Thus a library of data tapes was built up which could be called up and operated on very quickly.

(d) The ease with which programs can be changed saves time and hence money in the long run.

Obviously the digital techniques employed could not enable real-time evaluation of parameters to be made. However, the data could be collected at the time of the experiment so this was not much of a restriction.

The use of digital techniques also has the advantage that programs and data can be exchanged between workers more efficiently than is the case with analogue information and instruments. In the long run this should result in better correlation between all work in the field. However, as the work in Chapter 4 will show, even digital techniques can be difficult to handle and understand. The problems in digital analysis can be summarised as follows:

(a) Acquisition and quality of data.

(b) Pre-processing of data.

(c) Evaluation of parameters.

It is one of the objects of this thesis to obtain suitable methods for dealing with the digital data resulting from the measurement of surface topography in the most efficient way. Some of these points will emerge in the individual chapters. In the remaining part of this chapter emphasis will be given to the methods that have been adopted to process the data correctly. Also some essential detail of the programs that have been written for the various chapters will be given although fuller details including block diagrams and instructions will be left to Appendix 2.

3.4.2 Acquisition and quality of data

The acquisition of the data has been briefly dealt with in Section 3.4.1 where it was explained that punched paper tape was taken from the data logger with each ordinate being in the form of three decimal digits followed by an end of word character.

There is no value in trying to get resolution of each measurement greater than 10 bits because of the inherent noise level of the signal; this is due not necessarily to instrument noise, but to environmental mechanical or electrical noise. Where possible noise has been reduced by appropriate techniques such as supporting the Talysurf on a mechanical shock absorbing table. Unfortunately the use of the Datum Attachment on the Talysurf, figure 3-1, although removing possible errors due to the skid, makes the instrument more sensitive to extraneous mechanical vibration because of the increased mechanical loop between the pick-up datum and the workpiece. It is therefore necessary to use some care when employing the Datum Attachment.

The errors due to the limited resolution (i.e. 10 bits in this case) is called quantisation error (Watts 1961). It is not important from the point of view of the measurement of averaging type parameters like RMS or R_a, but it can be of importance when effects due to sampling and the definition of parameters are taken into account. This will be briefly explained later. For an RMS evaluation, for example, using Shepards correction if q is the resolution limit then the error in RMS is q $\sqrt{12}$ which is negligible in the 10 bit case.

In all this work equi-spaced samples have been taken for ease of instrumentation but this is not necessarily the best for any application, (Linden 1959). Second order sampling which uses overlapping trains of equi-spaced samples can be used in band-pass signals with a large reduction in the amount of data over first order sampling (equi-spaced data).

However, even the use of equi-spaced data has its problems as will be readily seen in Chapter 4 where it is shown that the values of many measured parameters depends crucially on the sampling interval.

3.4.3 Pre-processing of data: principles

This is one of the most important of the techniques that have been developed here to meet the need for getting useful information out of surface data. Basically the problem is as follows: The profile graph emerging from any stylus trace instrument can have two sets of extraneous information contained within it, the one usually long wavelength, and the other short wavelength. Both of these can be troublesome and can give rise to incorrect results. The long wavelength errors affect the average type of measurement such as autocorrelation, R_a etc., whereas the shorter wavelength errors affect the discriminating parameters such as the derivatives or curvatures on the surface. It must be conceded that both of these types of parameters are of fundamental importance in surface topography and are of particular importance in relation to the work in this thesis.

The nature of the long wavelength extraneous component is usually two-fold; one instrumental and the other natural. The long wavelength error introduced instrumentally is caused because of the imperfect levelling of the specimen relative to the mechanical datum of the instrument. This has the effect of producing a ramp-like profile signal on the Talysurf chart. Other effects are the general curvature or shape of the surface being measured or even the presence of waviness which is due to imperfect machining of the surface. Similar effects can occur in the digital record of the displacement of one body as it slides over another (the ride) because of errors in the friction apparatus.

The short wavelength (or sometimes better expressed as high frequency) extraneous components are usually due to electrical noise like the mains or vibration due to motors, gearboxes and general impulses in the vicinity of the instrument.

In both of these cases, both high and low frequency, some form of filtering technique must be adopted and, in general, the filtering is best done digitally because the characteristics and accuracy can be closely controlled.

Many forms of filtering are possible including the fitting of a least squares line or polynomial through the profile. The former removes the errors due to instrument but not other errors. Least squares polynomials like Legendre polynomials can be used, or even Chebychev polynomials which, although not least square, do find the minimum divergence between the profile and the polynomial. The disadvantages of all these is that some knowledge of the polynomial degree of the error must be known otherwise distortion can sometimes result. The best method is to use a true digital filter which does not require a knowledge of the profile. These will be explained in the next subsection.

A special filter has been devised to be used specifically in work on surface topography which can provide a high degree of accuracy and realism (Whitehouse 1967). This differs from the standard 2-CR filter digital technique derived earlier (Whitehouse and Reason 1965).

3.4.4 Pre-processing of data: Digital filtering

In general if y_N is the N^{th} ordinate of the profile and the ordinate spacing is uniform, then if the mean line found by the low pass filter is M_N for the N^{th} value then, in general, (Haykin and Carnegie 1970)

$$M_N = \sum_{i=0}^{N} \alpha_i y_{N-i} + \sum_{i=1}^{N} \beta_i M_{N-i} \qquad (3\text{-}1)$$

In other words the N^{th} output from the filter in digital form can be expressed in terms of all other preceding inputs and outputs.

If all the β_i's are zero then

$$M_N = \sum_{i=0}^{N} \alpha_i y_{N-i} \qquad (3\text{-}2)$$

where α_i are weighting factors found from the impulse response of the desired filter.

Where β_i are not zero then the filter is called recursive or closed loop. For example, a single-stage digital filter could be made by the expression

$$M_N = \alpha y_{N-1} + \beta M_{N-1} \qquad (3\text{-}3)$$

where the choice of α and β determine the position of the low cut break point and the gain of the transmission characteristics.

Because the choice of the β_i's are critical due to the closed loop nature of the digital filter the form of equation (3-2) was used.

Consider the standard instrumental method for removing the extraneous long wavelength components. This consists of two CR filters in cascade. The impulse response h(t) is given by

$$h(t) = \delta - (2 - t/RC) \exp(-t/RC) \,.\, 1/RC \qquad (3\text{-}4)$$

where δ is the unit impulse and RC is the time constant which can be written in terms of a high and low pass component $\overline{h}(t)$ and $\underline{h}(t)$ thus

$$\overline{h}(t) = \delta - \underline{h}(t) \qquad (3\text{-}5)$$

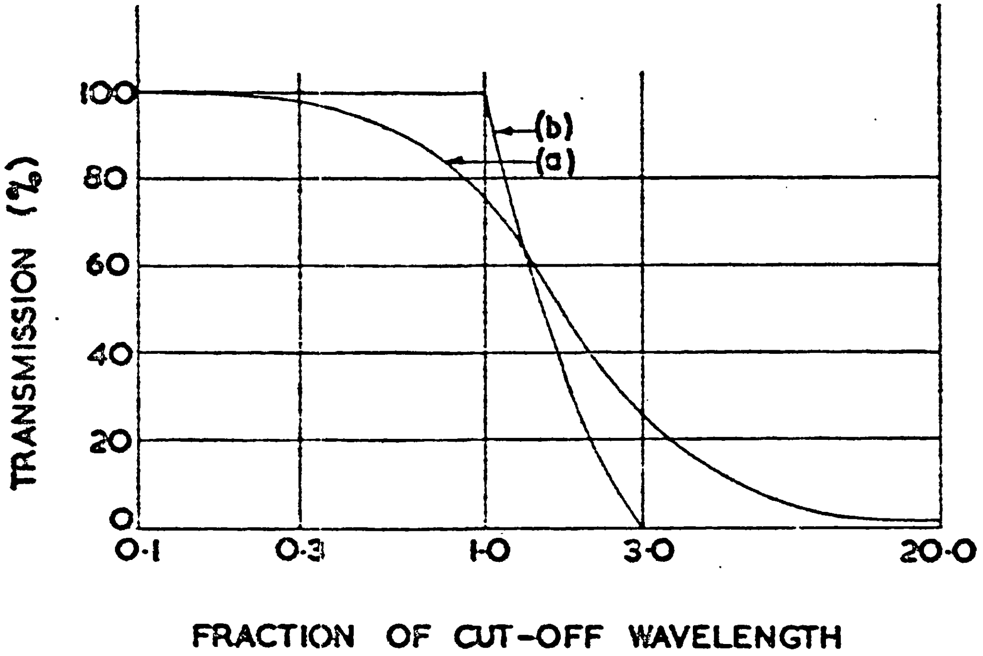
The true output from the filter g(t) is given by

$$\begin{aligned} g(t) &= \int_{-\infty}^{t} \overline{h}(t-\tau)\, y(\tau)\, d\tau \\ &= \int_{-\infty}^{t} \delta(t-\tau)\, y(\tau)\, d\tau - \int_{-\infty}^{t} \underline{h}(t-\tau)\, y(\tau)\, d\tau \qquad (3\text{-}6) \end{aligned}$$

= y(t) - m(t) where y(t) is the original profile and m(t) is the mean line at t. (3-7)

FIG 3-7

IMPROVED TYPE OF WAVEFILTER FOR USE IN SURFACE FINISH MEASUREMENT.



(a) STANDARD WAVEFILTER CHARACTERISTICS

(b) 3:1 PHASE – CORRECTED FILTER

Equation 3-7 can be rewritten non-dimensionally in terms of α the ratio of distance to the cut-off. Thus

$$g(\alpha) = y(\alpha) - m(\alpha) \qquad (3\text{-}8)$$

here (3-4) becomes $\bar{h}(\alpha) = \delta' - A\exp(-A\alpha).(2-A\alpha)$
$A = \lambda/\nu_p RC$ where λ is the cut-off, ν_p is the tracking speed of the pick-up, $\alpha = x/\alpha$ where x is the distance along the profile. (3-9)

Although the standard wavefilter has been useful in practice it has certain disadvantages when dealing with research problems on surface topography because of the following:

(a) The characteristic is such that the mean line is not smooth for high frequency profiles.

(b) The mean line is not flat up to the cut-off of the filter.

(c) There can be phase distortion of the filtered signal in some cases where the signal is near to the filter cut-off.

Because of these disadvantages and because of the research work required in this thesis and elsewhere a new filter was devised, it is called the phase-corrected filter. Figure 3-7 shows a comparison of the amplitude characteristic of the standard filter and the phase-corrected filter. A published paper on this work is given in Appendix 3 (Whitehouse 1968).

This phase-corrected filter is developed in the following way: No phase distortion of the filtered output implies that all components of the input waveform, that have not been completely rejected, are not shifted relative to each other in their passage through the filter. Strictly they should not be shifted in time at all. However, this is impossible, because if

$$H(\omega) = R(\omega) + jX(\omega) \qquad (3\text{-}10)$$

is the frequency characteristic of a filter, the phase angle ϕ is given by

$$\phi = \tan^{-1} X(\omega)/R(\omega) \qquad (3\text{-}11)$$

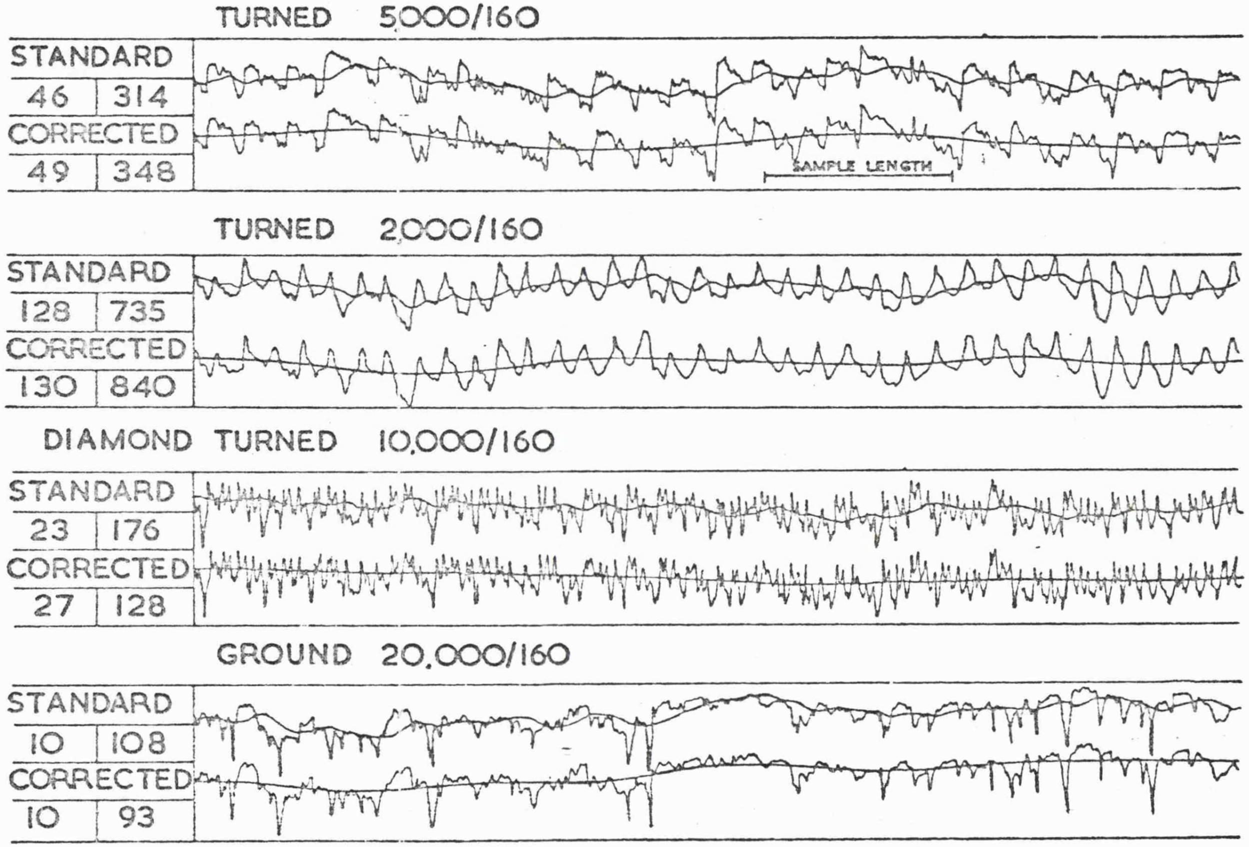
From which for ϕ to be zero $X(\omega) = 0$ i.e. $H(\omega)$ is real.

$H(\omega)$ can only be real if the impulse response is an even function, that is it extends equally on both sides of the time origin into both the future and the past which is impossible. However by shifting the axis of symmetry of the impulse response from the time origin to $t = t_0$ this introduces only a delay into the filter (which is equivalent to a linear phase term) and shifts all the components relative to each other by an equal amount, thus removing phase distortions. There are problems involved in having to shift the time origin a sufficient amount but these are explained elsewhere (Whitehouse 1968).

Under these conditions and taking the improved characteristics into account then equation (3-5) now becomes -

FIG. 3 –8

COMPARISON OF PHASE-CORRECTED AND STANDARD FILTER MEAN LINES.



$$\bar{h}(t) = \delta\,(t-t_0) - \underline{h}\,(|t-t_0|) \qquad (3\text{-}12)$$

and

$$g(t) = y(t-t_0) - \int_0^t \underline{h}(t-t_0-\tau)\; y(\tau)\,d\tau \qquad (3\text{-}13)$$

and the normalised impulse response of the desired characteristic.

$$\bar{h}(\alpha) = \delta'(\alpha-\bar{\alpha}) - \frac{1}{\pi^2(1-B)} \cdot \frac{\mathrm{Sin}\pi(1+B)(\alpha-\bar{\alpha})\mathrm{Sin}\pi(1-B)(\alpha-\bar{\alpha})}{(\alpha-\bar{\alpha})^2} \qquad (3\text{-}14)$$

where B is the ratio of wavelengths having unit to zero transmission and is equal to 1/3 in the chosen case. $\bar{\alpha}$ is the normalised version of t_0.

Equation 3-13 which incorporates equation 3-14 is easy to evaluate and operate in the computer. The values of α_i corresponding to the digital values of the low pass component of equation (3-14) are tabulated in the computer. Some examples showing how effective the new type of filter is are shown in figure 3-8.

Use of such a filter ensures that what may conveniently be called a realistic profile waveform will result (Whitehouse 1968). It must be emphasised that the real advantage of a digital filter over that of a polynomial fit under such circumstances is the predictability of its behaviour on any profile.

All practical waveforms measured throughout any of the experiments described in this work are first pre-processed by the digital filtering method described. All parameters to be measured are subsequently derived from the filtered output.

Short wavelength filtering can be used to remove noise by exactly the same method except that the component involving the impulse function is now no longer required. This means that the output is low-pass. Obviously, because short wavelengths are being removed the extent of the weighting function of the filter is very much shorter than that used to establish a mean line for the removal of long wavelength errors. The only occasion for the use of the high frequency filter in this work was in the analysis of friction waveforms.

3.5 Digital analysis

3.5.1 General principles

As in analogue techniques there is a considerable skill in digital analysis. Unfortunately many engineers skilled in analogue techniques do not realise that it is not obvious to go from one to the other. The measurement of derivatives is just one example. Most investigators of surface topography have used three-point analysis for the majority of the definitions of peaks and derivatives etc., as will be made clear in Chapter 4. However, there are better, although more laborious, ways of doing the job. For instance, there are techniques of numerical differentiation, integration, interpolation and extrapolation, all of which should be used if high accuracy is required. However, it must be admitted that these methods are not always necessary. For instance, in the evaluation of the mean line from the standard 2 CR filter the convolution integral is evaluated by using the trapezoidal rule. But it is not necessary to work out a mean line point for every profile point; one in every few profile points can be evaluated and a linear or parabolic interpolation made between the points. This saves computing time. In what follows it will be pointed out occasionally where numerical techniques are needed.

One fundamental point is that the numerical analysis formula to be used, the sampling rate, and the quantisation interval, are all intimately tied together in questions of accuracy. For example, in the three point definition for a peak (namely that the central ordinate should be highest), if the sampling rate is high compared to the bandwidth of the signal, and if the quantisation interval is large compared to the amplitude of the signal then not many peaks will be counted. But using another more comprehensive definition of a peak might increase the count. This could also be achieved in other ways, for instance by reducing the quantisation interval.

3.5.2 Evaluation of parameters

In the course of the following work a large number of parameters have had to be measured. For these a number of programs have been written. The following is a list of the parameters that have been necessary to evaluate:

(1) Autocorrelation functions and power spectra.

(2) Statistical distributions of various parameters including ordinate heights, peaks, valleys, curvatures of both for different heights, slopes and second derivatives, both filtered and unfiltered. These distributions are computed together with the necessary moments and extremes of the distributions.

Additional programs have been written to measure particular points, e.g.

(a) The evaluation of the envelope of a circular body having one degree of freedom moving across the profile waveform.

(b) The generation of random profiles according to various statistics.

Other programs distinct from these have had to be written to either verify or work out numerically some of the theoretical formulae. These include the evaluation of envelope behaviour, the distributions of peaks and valley curvatures and height distributions etc. These will be outlined in Appendix 2.

In all the programs involving measurements on data acquired during experiments such as those obtained from Talysurf profiles of surfaces or from friction and experiments are initially transcribed from paper tape onto magnetic tape. They then become part of a library of data tapes, each track or profile being classified by a number and a magnetic tape name. Preceding each set of data on the magnetic tape is an indentifier containing information on magnifications, manufacturing process, data etc. When using this data in any of the programs the identifier is automatically printed at the top of the line printer page.

The advantage of transcription onto magnetic tape (or disc) is that the program can be written more efficiently, repeated scanning of the data being possible.

All programs are written in I.C.L. version of Fortran IV.

3.5.3 Autocorrelation and power spectra

The formula used for the evaluation of the autocorrelation is given by C(β) where

$$C(\beta) = \overline{y(x)y(x+\beta)} \qquad (3\text{-}15)$$

or for a practical record

$$C(\beta) = \frac{1}{L-\beta} \int_{0}^{L-\beta} y(x) \; . \; y(x+\beta)dx \qquad (3\text{-}16)$$

where L is the length of the record and y is the profile of zero mean value. In the program, equation (3-16) is divided by the variance to give the normalised autocorrelation. In equation (3-15) which is obtained from a single profile record the ergodic principle is assumed (Lee 1960) that is, the time (or space) average as given in the equation is equivalent to the ensemble average.

If T is the spacing, M is the lag number and N is the Number of ordinates in the record, then equation (3-15) becomes in the normalised and digital form.

$$C(MT) = \frac{1}{N-M} \sum_{i=1}^{N-M} y(iT) \, . \, y(iT+MT) \Big/ \frac{1}{N} \sum_{i=1}^{N} y(iT)^2 \qquad (3\text{-}17)$$

The structure function S (β) is given by

$$\overline{(y(x)-y(x+\beta))^2} \qquad (3\text{-}18)$$

which is often more reliable for large values of β because it effectively removes some elements of drift in the mean value. In the program S(β) is evaluated at the same time as the autocorrelation function.

Notice that $S(\beta) \;=\; 2(C(0)-C(\beta))$ (3-19)

indicating that there is no difference in the information if the data are strictly stationary.

The power spectrum, or more correctly the power spectral density, is given by (see for example Bendat and Piersol 1966).

$$P(f) = 2 \int_0^\infty C(\beta)\, W(\beta)\, \mathrm{Cos} 2\pi f\beta \; d\beta \qquad (3\text{-}20)$$

where W(β) is a lag window used for reducing the presence of misleading information introduced into the power spectrum because of the abrupt truncation of the autocorrelation function at the maximum allowable value of β (β_{max}) deemed suitable for reliability. In practice β_{max} is usually about 10% of L.

$$W(\beta) = 0.5 + 0.5\mathrm{Cos}(\pi\beta/\beta_{max}) \qquad (3\text{-}21)$$

The form of $W(\beta)$ here used is due to Hanning (Blackman and Tukey 1958). There are more direct ways of getting the power spectrum than via the autocorrelation function i.e. direct from the signal itself, but it was considered best to proceed in the way indicated because of the additional need for the autocorrelation function itself.

3.5.4 Other parameters

Derivatives, curvatures etc. In the investigation into the three-point analysis technique discussed in Chapter 4 the definition of a first derivative was taken digitally to be

$$y_0' = \frac{y_{+1} - y_{-1}}{2T} \qquad (3\text{-}22)$$

where T is the ordinate spacing. A more accurate formula would use more than the three ordinates; for example

$$y_0' = \frac{1}{60T}\left[y_3 - 9y_2 + 45y_1 - 45y_{-1} + 9y_{-2} - y_{-3}\right] \qquad (3\text{-}23)$$

In cases other than testing three-point analysis work this is preferred. Similarly for the second differential instead of

$$y_0'' = \frac{(y_{-1}-2y_0-y_{-1})}{T^2} \qquad (3\text{-}23)$$

it, more accurately, should be (HMSO 1956)

$$y_0'' = \frac{1}{180T^2}\ (2y_3-27y_2+270y_1-490y_0+270y_{-1}-27y_{-2}+2y_{-3}) \qquad (3\text{-}24)$$

In working curvatures out use is made of the formula

$$\frac{1}{R} = \frac{y_0''}{(1+(y_0')^2)^{3/2}} \qquad (3\text{-}25)$$

which reduces in the region of peaks and valleys where the interest usually lies to

$$\frac{1}{R} \sim y_0''$$

because in these regions $y_0' \sim 0$ for a peak y_0'' is negative and for a valley it is positive.

Once the distributions of these parameters have been found it is a simple matter to work out the basic moments of them. This is the same procedure for any distribution. For example if the probability density of the profile height is f(y).

The mean value $\bar{y}$ is $\int_{-\infty}^{\infty} yf(y)dy$ (3-26)

The average $\int_{-\infty}^{\infty} |(y-\bar{y})|f(y)\ dy$ (3-27)

The variance is $\int_{-\infty}^{\infty} (y-\bar{y})^2.f(y)\ dy$ (3-28)

$= (\bar{y})^2 = \sigma^2$ (the RMS value squared)

The skew is given by

$$\frac{1}{\sigma^3} \int_{-\infty}^{\infty} (y-\bar{y})^3 \, f(y) \, dy \qquad (3\text{-}29)$$

and the kurtosis or excess by

$$\frac{1}{\sigma^4} \int_{-\infty}^{\infty} (y-\bar{y})^4 \, f(y) \, dy - 3 \qquad (3\text{-}30)$$

Other moments could be measured but in practice these are unreliable because any extreme freak peak ordinate can dominate the results.

Other parameters based on derivatives can be measured, one of these is the total length of curve.

This is given by

$$\text{Total length} = \frac{1}{x_2-x_1} \int_{x_1}^{x_2} \sqrt{1+(y_0')^2} \quad dx \qquad (3\text{-}31)$$

This works out as the running secant of the surface. Hence

$$\text{RMS secant} = \sqrt{1+ \text{mean square slope}} \qquad (3\text{-}32)$$

One more very important parameter that is measured is the average wavelength parameter which was initially proposed by Mr. R. C. Spragg. This has been defined via the mean square angular frequency $\overline{\omega^2}$.

$$\overline{\omega^2} = \frac{\int_0^\infty \omega^2 P(\omega)\, d\omega}{\int_0^\infty P(\omega)\, d\omega} \qquad (3\text{-}33)$$

$$= \overline{(y')^2} \;/\; \overline{(y)^2} \qquad (3\text{-}34)$$

The average wavelength is defined from equation (3-34) by inversion to give

$$\lambda_{RMS} = 2\pi \,.\, \frac{y_{RMS}}{y'_{RMS}} \qquad (3\text{-}35)$$

which is a general expression for either a random or periodic waveform. See Spragg and Whitehouse (1971) for a discussion of this parameter.

In the program both average and RMS values are worked out. The bearing ratio, which is defined from the amplitude density function, is also worked out. It is in effect simply unity minus the amplitude distribution function i.e.

$$BR(y_1) = \int_{y_1}^{\infty} f(y)\, dy \qquad (3\text{-}36)$$

and has been used extensively in surface typology.

3.5.5 Locus of movement of one cylinder on another in the crossed cylinders-machine

Essentially this involves the working out of the path that a smooth upper cylinder would take in running over a rough lower cylinder.

Basically in 2D the method consists of plotting the path of a circle of known radius across the peaks of the profile. It is essentially a mechanical filtering device. The computer technique, as in conventional method of digital filtering, is straightforward but tedious. First the circle radius R has to be modified into an ellipse to take account of the magnification differences between horizontal V_h and vertical magnifications V_v of the data:

$$\text{i.e.}\quad y = RV_v\ (1 - \sqrt{1 - (\frac{x}{RV_h})^2}\ ) \qquad (3\text{-}37)$$

Let the values of y for equal increments of x be $c_{-L}\ \ldots c_0 \ldots c_L$ where c_0 is the point on the circle corresponding here to $x = 0$.

To find the position of the point of contact of the envelope at any point on the profile, say a_k, the ellipse ordinates are positioned such that c_0 is lined up coincident with a_k. The two series of ordinates are then added up. If the maximum sum is at a position corresponding to a_s where s is the number of ordinates spacings measured from the position k then the height of the envelope at position k is $a_s-(c_0-c_s)$ and the difference between the envelope and a_k is simply $a_s-(c_0-c_1)-a_k$. This operation is repeated for each profile ordinate in turn along the available length. The R_p value is then

$$R_p = \frac{1}{N-(2L+1)} \sum_{k=L}^{k=N-(L+1)} ((a_s - a_k) - (c_0 - c_s)) \qquad (3\text{-}38)$$

where s takes a different value for each k.

Obviously the length of computation here depends mainly on the time for finding the position of contact at each position of the circle relative to the profile and this in turn depends on the extent of the ellipse.

Choosing the extent of the ellipse for each radius and magnification ratio is a matter of compromise. Some limit has to be imposed even if only for reducing computing time. A normal criterion is to limit the vertical depth of the arc to a given fraction of the chart. For instance on a Talysurf this might be 25%. Even under this restriction the extent of the scan required can be large, for example if the ordinate spacing is 2.5 μm and the radius is 50 mm then the extent of the envelope can be over 1,300 ordinate positions if the magnification is low, say 500X.

This procedure can be extended to take into account a rough upper cylinder, in which case $c_L \ldots c_0 \ldots c_L$ are not simply ellipse ordinates they are magnified ordinates taken from a cross-section of the upper cylinder. Strictly this cross-section would have to be continually changed but for most cases it is not necessary.

3.6 Some details of a program

To bring together some of the topics discussed in sections 3.4 and 3.5 it will perhaps be constructive to examine a typical program. Here no attempt will be made to list the full program, but some details and a flow diagram,are given in Appendix 2 where some details of other programs that have been written will be given.

The program here selected is called PROF. It works out many parameters associated with surface texture (or friction or ride). Since being originally written by the author it has had some small changes made to it at various times, particularly in the format arrangements by my colleagues Messrs. A. Bykat, G. Burger and D. Kinsey. The technical content of this program and the digital techniques presented are, however, virtually unchanged from the original version.

The program can be split up into input routines, reference lines and parameters and this sub-division will be followed below.

3.6.1 Input routine

The data emerging from the data logger consists of words of three decimal digits followed by an end of work character. In the course of the work two different data logging systems were used, one of five channel and the other of eight channel; therefore the actual width of the paper tape used could correspond to either of these situations and the format of the character information depended on the data logger used.

For speed of computation the paper tape information was transcribed onto computer magnetic tape by means of an editing routine called RSFE which, as the initials imply, is a surface finish editor. This puts the paper tape information onto magnetic tape in the following form:

(a) An identifier comprising of up to 80 characters.

(b) The data in blocks of one thousand ordinates.

(c) A number giving the total number of ordinates on the tape.

Another advantage of transcription apart from that of being able to rescan the data ordinates is that in this editing routine checks on the data can be made. This saves time and money when the main large program is read in.

The editing routine also enables some degree of organisation of the surface profiles on magnetic tape. After transcription, during which time the ordinates were automatically listed on the line printer, a condensed list of the total number of profiles on the magnetic tape is printed together with their identifiers and number of ordinates. No other details of the editor will be given here.

To call data from the magnetic tape two subroutines

SEEKSURFACE and READSURFACE are used (both FORTRAN and PLAN versions are available). SEEKSURFACE uses as arguments the number of the surface profile on the magnetic tape and the name of the magnetic tape. It positions the magnetic tape at the beginning of the required surface and inputs into the allocated store locations both the profile identifier and the number of ordinates in the profile.

READSURFACE reads in the data from the profile in blocks of 1000 putting them into prescribed locations in the store. It makes available the number of actual ordinates that have been read-in, the last block for instance will rarely contain the full 1000 ordinates. As an example of how this works the statement CALL READSURFACE (NN, Y(4001)) puts 1000 ordinates in the Y array starting at 4001 and it puts the number of ordinates with value other than zero that have been read in the block into NN.

Having the number of profile ordinates available on the magnetic tape enables a prescribed length of surface profile to be read in. This is useful when checking R_a or peak values which have been taken using an analogue surface measuring instrument because in the Talysurf, for instance, the meter assessment does not simply start at the beginning of the trace, it depends on the meter cut-off. Thus for the 0.25 mm cut-off, it is closer to the end of the traverse than it is for the 2.5 mm cut-off. Knowing the number of ordinates before the start of the run also ensures that no time is wasted. In the program the number of ordinates that can be skipped over at the start is called IGNORE.

Other useful features of the input routine are that the control variables are coded, enabling them to be put in any order. Also the control data for up to ten profiles can be read-in in one go. This does not mean that ten profiles have to be worked on every time. Simply putting TRAV (the assesment length) equal to zero indicates that the end of the batch has been reached.

Other points in this programme are as follows: (a) The input control variables are examined in a subroutine CHECKS. If an error is detected or a questionable control value the program either adjusts it to a value which is reasonable or it flags an error on the line printer. (b) For the purposes of display the routine for plotting results on the line printer automatically scales the values to take best advantage of the width of the paper.

Another useful feature of the programme is in the final output. This is a listing of the major parameters that have been obtained from all the surfaces that have been run in the batch. This listing covers usually all the details of the distributions etc. However, it has not been possible to include in the listing any information about the autocorrelation function because it is usually so complicated as to need visual assessment.

3.6.2 Processing

Turning to the process of digital filtering the program allows quite a variation in the method of operation. Either the phase-corrected or the standard 2 CR filter can be used or neither. One of the first steps in the program is to set the weighting function into store. For the phase-corrected filter this is done in the following way using ICL FORTRAN IV.

Where

AA(500) is the array used in this instance to store the function

B is the drop off rate required

L is the number of weighting factors (assumed odd)

SUM is the store used for a normalising factor

C is the number of weighting factors per cut-off

PI = 3.14159

K is the ratio of ordinates to weighting factors.

The program reads:

```
5008    AA(250)    = 1+B - the value of the central weighting factor
        SUM        = 1+B
        DO 5009    I = 1, L/2
        ALPI       = (1+B) *PI *I/C
        ALP2       = (1-B) *PI *I/C
        AA(250+I)  = SIN (ALP 1)* SIN (ALP2) / ((ALP 2**2) / (1-B))
```

Using the formula described in Section 3.4.4

```
       AA(250-I) = AA(250+I)       :  The weighting function is
                                      symmetrical so only half needs
                                      to be calculated.
5009   SUM = 2 * AA(250+I) + SUM   :  The weighting function
       FACTOR = SUM /C                normalising factor
       DO 5010 I = 250 - L/2, 250 + L/2
```

```
5010  AA(I) = AA(I)/FACTOR     : The area in the weighting function
                                 is normalised.

      J = K*(L-1) +1           : This works out the number of profile
                                 ordinates covered by the weighting
                                 function.

      ITRAK = ITRAV - J+1      : This gives the number of ordinates
                                 in the assessment.
```

After putting the weighting function into store the actual convolution operation to get the mean line has to be carried out. This convolution operation is particularly simple when carried out in the computer but it can be time consuming. The following is an example of how it is done:

```
     DO    5011    KI = 1, ITRAK
     SUM  =  0.0
     DO    5012    I = 0, L-1
     SUM  =  SUM + AA (250 - L/2 +I) * Y (I*K + KI)
5012 CONTINUE
     Y(KI) = Y(L * K/2 + KI) - SUM/C
```

Here the mean line point at Y(L * K/2 + KI) is contained in SUM. This is taken from the profile value which is Y(L * K/2 + KI) which is then shifted to position KI in the Y array for clarity later on.

In this particular program a mean line point is worked out for every available point but in some of the other programs this is not necessarily done; linear interpolation is used to estimate the mean line between computed mean line points. In all the programs referred to in Appendix 2 one of the special features has been the options built in to allow many of the features to be utilised or not as required.

Another, perhaps, more important feature in these programs has been the continual use of the line printer for a simple pictorial display of evaluated results. The line printer is far better than a graph plotter for simple display purposes, because of its much greater speed. The author believes firmly in the importance of a visual display in addition to numerical information for most computer applications, to give impact. All the programs described make great use of this feature especially in the presentation of distributions, mean lines, spectra etc.

An example of how this plot routine works will be given in the following section.

3.6.3 Evaluation of parameters

As a simple example of how the parameters are evaluated and displayed consider the section of the program relating to the distribution of the filtered slope.

The value of the input variable JOHN determines whether or not to apply any high cut filtering. If this is required it is carried out in exactly the same way as for the low cut filtering described in the previous section, except that now the desired profile IS the mean line that has resulted from the convolution operation. It will then be stored in the Y array. The differential is then given in the following routine:

If V(M) is the magnification vertically

PS(M) is the profile ordinate spacing

AB(I) is the array in which the distribution of slope values is stored

Then D1 = 1.0 /30000.0 / V(M) /PS(M)

```
     SLOMAX  =  -10000.0  )
                          )  These set artificially high and
     SLOMIN  =   10000.0  )  low limits for the slope.

     ITRAK   =   ITRAK-6

     DO 1801  IAT = 1, ITRAK, 1.

     Y(IAT)  =   (Y(IAT+6)- 9*Y(IAT+5) + 45*Y(IAT+4) -45*Y(IAT+2)
                 +9* Y(IAT+1) -Y(IAT))
                              At this stage only differences
                              are used they are scaled later on.

     SLOMAX  =  AMAXI (SLOMAX, Y(IAT))  )  These routines determine
                                        )  the highest and lowest
     SLOMIN  =  AMINI (SLOMIN, Y(IAT))  )  slopes.

1801 CONTINUE

     DELTA 2 =  (SLOMAX - SLOMIN) /41    : This determines the
                                           slope distribution
                                           interval.

     DO 1802 IAS = 1, 41

     AB (IAS) = 0.0                      : This clears the array.

1802 CONTINUE

     DO 1803 IAS = 1, ITRAK, 1

     KAT = N INT(Y (IAS) /DELTA 2) + 21  : Determines location
                                           adds to the count

1803 CONTINUE
```

The next step calculates the moment of the distribution.

```
     SUM 1 = 0.0

     SUM 2 = 0.0
     SUM 3 = 0.0
     SUM 4 = 0.0
     SUM 5 = 0.0
     DO 1804 IAS = 1, 41.
1804 SUM 1 = SUM 1 + AB(IAS) * (FLOAT (IAS) - 0.5)
     SUM 1 = SUM 1 /ITRAK          : Works out mean value
     DO 1805 IAS = 1, 41
     SUM 2 = SUM 2 + ABS (FLOAT (IAS) - 0.5 - SUM 1) * AB(IAS)
     SUM 3 = SUM 3 + (FLOAT (IAS) - 0.5 - SUM 1) **2*AB(IAS)
     SUM 4 = SUM 4 + (FLOAT (IAS) - 0.5 - SUM 1) **3*AB(IAS)
     SUM 5 = SUM 5 + (FLOAT (IAS) - 0.5 - SUM 1) **4*AB(IAS)
1805 CONTINUE
```

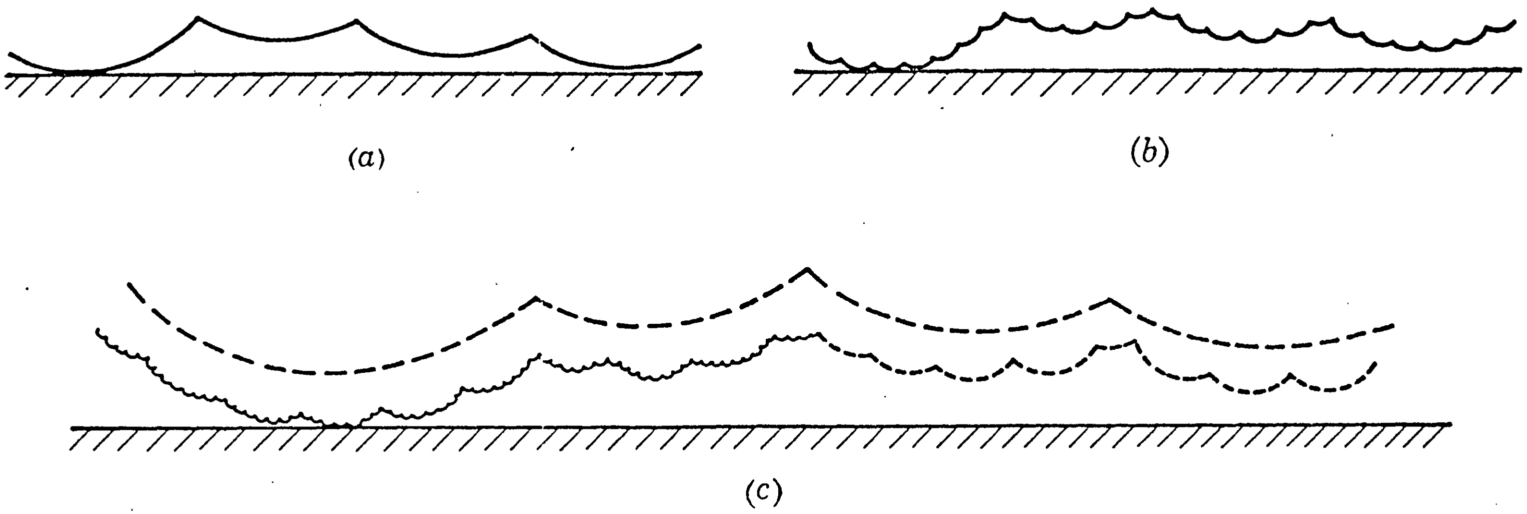
It is, of course, possible to work out the moments from the slope values direct but this, although more accurate, takes more time. The moments are subsequently converted to average, RMS, skew and excess values. Also the ordinate differences in y(IAS) are changed to real parameters like slope in degrees etc. When the results are ready a routine called PLOT is used which enables the line printer to be used as a simple plotter. First the maximum frequency in the distribution is found. This is called PLAX.
Also part of the array AB is used to store the character information required for the routine. The locations 42 to 46 are used.
VARY 2 is the value of the slope, VARY 1 is the tangent and IKIP is the character number representing the frequency at slope value VARY 2. Thus

```
     CALL  PLOT (64,40,40,AB(42))    : Clears the array AB from
                                       42  onwards.

     CALL  PLOT (26,IKIP,40,AB(42)) : Puts an asterisk in the
                                       location in the AB array
                                       corresponding to the value
                                       IKIP.

     WRITE (2,1809) VARY 2, VARY 1, (AB(L), L = 42,46)

1809 FORMAT (1X, F10.4, 2X, F10.4, 6X, 5A8)
```

This routine puts the slope value, tangent, and an asterisk IKIP locations along the line printer width starting from the end of the field of VARY 1.

Obviously to cover all cases of the evaluation of parameters covered in the entirety of the programs would require a great deal of space. These selected examples are meant simply to give an idea of the methods adopted.

Figure 4-1. Models of surfaces containing asperities of differing scales of sizes. When the deformation is elastic the relationships between the area of contact (A) and the load (W) are as follows: (a) $A \propto W^{\frac{4}{5}}$; (b) $A \propto W^{\frac{14}{15}}$; (c) $A \propto W^{\frac{44}{45}}$.



4. THE RANDOM SIGNAL MODEL OF A SURFACE PROFILE AND ITS ANALYSIS IN RELATION TO SURFACE CONTACT

4.1 Introduction

It is clear from the review of Chapter 2 and Appendix 1 that surface finish is most important in the functional behaviour of surfaces used in common engineering practice; a particular area where surface finish is highly significant involves those applications concerned with surface contact. In the past the methods used in the specification of the geometric features of surfaces have been inadequate. They have been expressed only in terms of the height of the profile and have not taken into account the spacing of the asperities or the differing scales of size present in the surface structure (see, for example, figure 4.1). At the same time those concerned with theories of surface contact have used theoretical models in which the surface is represented as an assembly of asperities. To a large extent the form of these models has been based upon theoretical convenience rather than an examination of the structure of surfaces met in practice. To a limited extent the divergence between theoretical models and the analysis of surfaces has been removed by the work of Greenwood and Williamson (1966) who used digital analysis of surface profiles as the basis for their asperity model. Nevertheless the overriding need, at the present time, is for an analysis of the problems involved in the function of surface contact based upon knowledge of the topography of surfaces used in engineering practice.

The work of Greenwood and Williamson although representing a major advance is still far from a complete or accurate representation of random surfaces such as those analysed in their work. In this chapter we shall take as our starting point a description of the profile as a random signal; it will then be shown how it is possible, taking due account of the digital techniques usually employed, to deduce theoretically those properties of the profile relevant in surface contact.

4.2 The model

4.2.1 General

To regard the profile of a surface obtained from a stylus tracer instrument as a random signal is, in a great many instances, not an unreasonable one. The chart recording obtained from say a ground or shot blast surface could equally well have been obtained from an instrument which measures wind gusts or many other widely differing physical quantities, for instance, in oceanography (Longuet-Higgins 1957), seismology (Liu 1968) and medicine (Krendel 1959). That this random waveform description of a manufactured surface is not illusory can be demonstrated from the autocorrelation function and height distributions of many such surfaces. In order to understand better how the random waveforms concept can be applied to problems in surface topography it is necessary to examine some of the statistical properties of random waveforms.

Any random waveform can be represented in more and more detail by consideration of higher orders of joint probability function, (Bendat 1958). If values taken from this process are

y_1, y_2, y_3 at x_1, x_2 and x_3 then $f(y_1, y_2, y_3; x_1, x_2, x_3)$ more completely defines the process than does $f(y_1, y_2; x_1, x_2)$. For most random signals, however, the second order joint probability density function adequately defines the statistics of the process (Lee 1960) For stationary processes $f(y_1, y_2; x_1, x_2)$ can be written $f(y_1, y_2; \beta)$ where β is simply the distance $x_1 - x_2$. If it is understood in what follows that the y_2 value is taken at a distance of β from y_1, then for convenience the joint probability density can be written $f(y_1, y_2)$.

The ordinate height probability density function $f(y_1)$ and the autocorrelation function $C(\beta)$ can both be obtained from the second order function $f(y_1, y_2)$ because $f(y_1)$ is given by

$$f(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\, dy_2 \qquad (4\text{-}1)$$

i.e. $f(y_1)$ is a marginal distribution function of $f(y_1,y_2)$,also the autocorrelation function $C(\beta)$ can be obtained because as an ensemble average $C(\beta)$ is given as a joint moment by

$$C(\beta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1 y_2\, f(y_1, y_2)\, dy_1\, dy_2 \qquad (4.2)$$

In the case of Gaussian or near Gaussian processes the second and higher order joint probability density functions are completely determined by the standard deviation of $f(y)$ and the autocorrelation function. In what follows we will always take the normalised form of the autocorrelation function unless stated.

Consider the joint probability density function of random variables y_1, y_2 .....y_N taken from a Gaussian random process. Assume that they have zero mean and unit variance then use can be made of the multi-normal distribution well known in statistics (Cramer 1946). This is given in terms of y_1, y_2.....y_N as follows:

$$f(y_1, y_2 \ldots y_N) = \frac{1}{(2\pi)^{N/2}|M|^{1/2}} \exp\left[\frac{-\sum\limits_{i,j=1}^{N} M_{ij}y_iy_j}{2|M|}\right] \qquad (4\text{-}3)$$

where $|M|$ is the determinant of M; M is given by the square matrix

$$M = \begin{pmatrix} d_{11} \cdots\cdots & d_{1N} \\ d_{21} & \vdots \\ \vdots & \vdots \\ d_{N1} & d_{NN} \end{pmatrix}$$

d_{ij} being the second moment of the variables y_iy_j and M_{ij} is the cofactor of d_{ij} in M.

Take for example, the joint probability density of two ordinates y_{-1} and y_0 from a Gaussian waveform with a correlation of ρ then

$$f(y_{-1}, y_0) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[\frac{\sum\limits_{i,j=1}^{2} M_{ij}y_iy_j}{2(1-\rho^2)}\right] \qquad (4\text{-}4)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left[\frac{1}{2(1-\rho^2)} \quad (M_{11}y_{-1}^{2}+M_{12}y_{-1}y_0+M_{21}y_{-1}y_0+M_{22}y_0^2) \right] \qquad (4\text{-}5)$$

from which

$$f(y_{-1},y_0) = \frac{1}{\sqrt{2\pi}} \exp\,(-y_0^2/2) \;.\; \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp \left[- \frac{(y_{-1}-\rho y_0)^2}{2(1-\rho^2)} \right] \qquad (4\text{-}6)$$

which can be written

$$f(y_0,\, y_{-1}) \;=\; f(y_0) \;.\; f(y_{-1}/y_o) \qquad (4\text{-}7)$$

where

$$f(y_0) = \frac{1}{\sqrt{2\pi}} \quad \exp\,(-y_0^2/2)$$

and

$f(y_{-1}/y_0)$ the conditional joint probability density function* of y_{-1} given first y_0 is

$$f(y_{-1}/y_0) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp \left[- \frac{(y_{-1}-\rho y_0)^2}{2(1-\rho^2)} \right] \qquad (4\text{-}8)$$

which is a Gaussian distribution of mean value ρy_0 and variance $(1-\rho^2)$.

*As an example of the relevance of these concepts to the subject of this thesis, in figure 4-10 the conditional probability density function obtained from the digital analysis of a practical surface can be seen.

Similarly for three ordinates having correlations of ρ_1 between adjacent ordinates and ρ_2 between the extreme ordinates:

$$f(y_{-1},y_0,y_{+1}) = f(y_0) \cdot (f(y_{-1}/y_0) \cdot f(y_{+1}/y_0,y_{-1}) \qquad (4\text{-}9)$$

where

$$f(y_0) = \frac{1}{\sqrt{2\pi}} \exp\,(-y_0^2/2)$$

$$f(y_{-1}/y_0) = \frac{1}{\sqrt{2\pi(1-\rho_1^2)}} \exp\left[-\frac{(y_{-1}-\rho_1 y_0)^2}{2(1-\rho_1^2)}\right]$$

and (4-10)

$$f(y_{+1}/y_0,y_{-1}) = \frac{(1-\rho_1^2)^{1/2}}{\sqrt{2\pi(1-\rho_2)\ (1+\rho_2-2\rho_1^2)}}$$

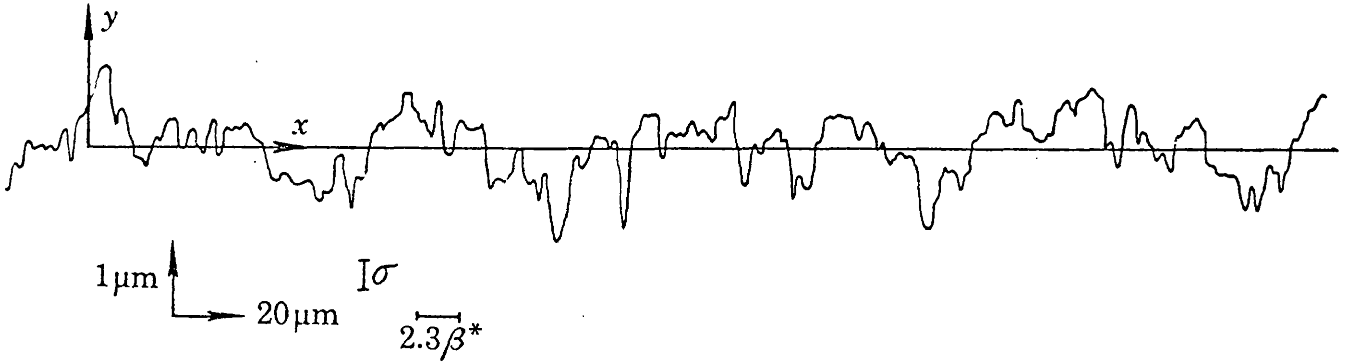
$$\exp\left[-\frac{(y_{+1}(1-\rho_1^2)-y_0\rho_1(1-\rho_2)-y_{-1}(\rho_1^2-\rho_2))^2}{2(1-\rho_1^2)(1-\rho_2)(1+\rho_2-2\rho_1^2)}\right]$$

The last term represents the probability density that y_{+1} occurs given that y_0 and y_{-1} have occurred.

From this it is seen that the second and all higher density functions can be specified completely in terms of the normalised autocorrelation function and the value of the variance of the signal (or more conveniently the RMS value).

Therefore it is proposed in this thesis that a completely adequate way of representing a random type signal such as is often met with in surface finish waveforms is to classify the surface in terms of its ordinate height distribution and normalised autocorrelation function.

Figure 4-2. Surface profiles of Aachen 64-13 showing co-ordinate system. The profile is of a surface chosen for a detailed analysis which is described later in the paper. The magnitude of the RMS value of the height distribution (σ) and the correlation distance β* are shown for comparison with the profile.



In particular, in this investigation it will be assumed that such a representation of the random surface is often best achieved by considering a Gaussian height distribution and an exponential autocorrelation. It will suffice here to say that the model can be justified in a number of ways.

First, and foremost, the author has found that a great number of manufactured surfaces have a height distribution and autocorrelation function which fit, or are a close approximation to, this model. Second, this model has been used in the past for various other problems, for instance in the scattering of electromagnetic waves from surfaces. Third, use of this model simplifies the mathematics sufficiently to allow some equations to be evaluated and hence conclusions to be drawn from them. Finally, further justification for such a model will be given in Chapter 7 with practical and theoretical examples.

Using a Gaussian height distribution and an exponential autocorrelation function means that specification of the model of the random waveform reduces to the use of two numbers only, the standard deviation of the distribution and the exponent of the normalised autocorrelation function.

The system of co-ordinates is shown in figure 4.2, the mean line through the profile will be taken as $y = 0$. In practice the signal will have zero mean because of the low-cut filter, mentioned in Chapter 3, which does not substantially affect the autocorrelation function. The probability of finding an ordinate at a height between h and h+δh is f(h)dh which for a Gaussian height distribution is

$$\frac{1}{\sqrt{2\pi}} \exp \ (-\tfrac{1}{2}y^2)dh \qquad (4\text{-}11)$$

Here the heights have been expressed in a normalised form $y = h/\sigma$ where σ is the RMS of the surface or the square root of the origin of the autocorrelation function when not normalised

The autocorrelation function normalised relative to σ^2 is given by

$$C(\beta) \ = \ \lim_{L\to\infty} \frac{1}{L} \int_{-L/2}^{L/2} y(x) \ . \ y(x+\beta) \ dx \qquad (4\text{-}12)$$

and is assumed to be exponential i.e.

$$= \ \exp \ (-\beta/\beta^*) \qquad (4\text{-}13)$$

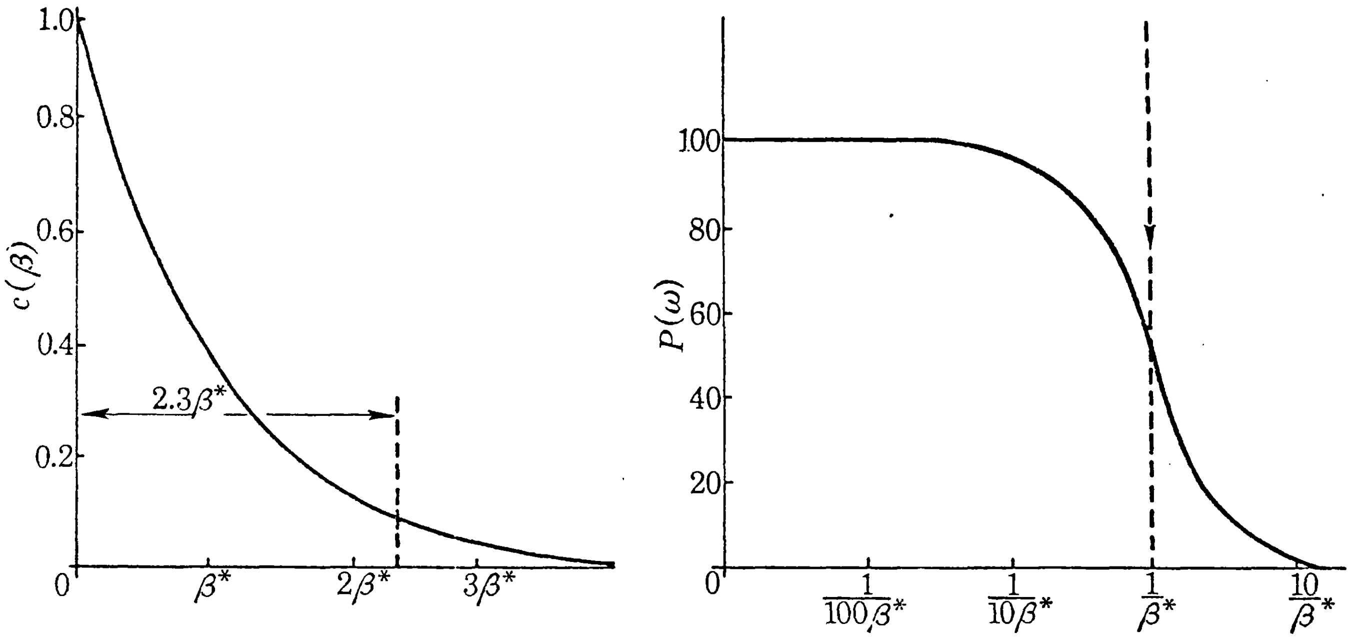
where β^* will be called the correlation distance*.

As the spacing between points on the profile is increased their heights become statistically less dependent on each other and β^* is a scaling factor indicating the rate at which this dependence (expressed by $C(\beta)$) declines towards zero.

For an exponential autocorrelation function the power spectrum

* We use the term 'correlation distance' to mark a distinction between ourselves and Peklenik (1967-8) who uses the term 'correlation length' for $2.3\beta^*$. The only reason for this distinction is that it is usual to specify a first order system in terms of the cut-off.

Figure 4-3. The model; autocorrelation function and power spectral density.



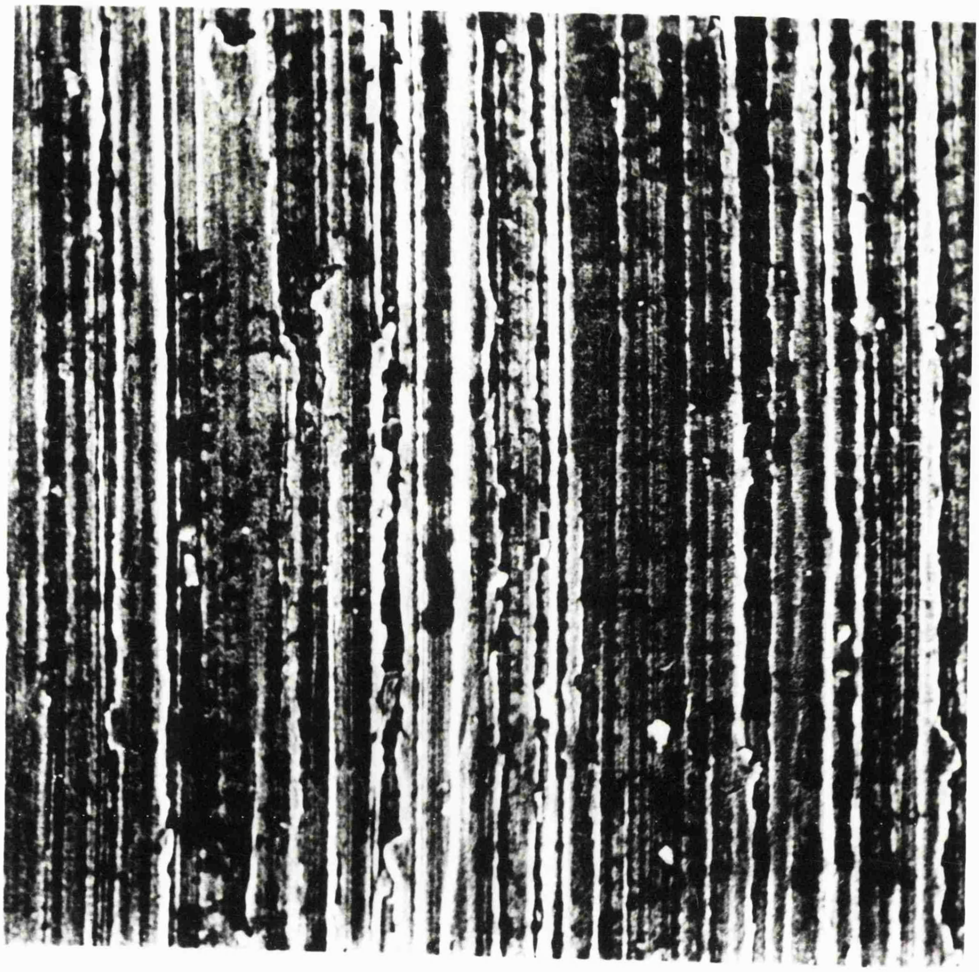


Figure 4-4. Micrograph of typical ground surface Aachen 64-13. (Magnification 1000X).

is represented by white noise limited only in the upper frequencies by a cut-off of 6db per octave (figure 4-3); this has the physical meaning that the main components of the profile lie within a band covering the lower frequencies (longer wavelengths). Shorter wavelengths do exist whose magnitudes decline in a way such that their amplitude is proportional to the wavelength. Figure 4-4 is a micrograph of a typical random surface which shows that these shorter wavelength components do, indeed, exist upon such surfaces.

4.2.2 Representation as a Markov process

Another important feature of the use of this particular model (and, in particular, the exponential correlation function) is that the surface can now be represented as a first-order Markov sequence; a point considered in more detail in the next chapter.

Putting $\rho_1^2 = \rho_2$ into equation (4-10), because of the property of exponentials, yields

$$f(y_{-1}, y_0, y_{+1}) = \frac{1}{\sqrt{2\pi}} \exp\,(-y_0^2/2)\, \frac{1}{\sqrt{2\pi(1-\rho^2)}}$$

$$\exp\left[- \frac{(y_{-1}-\rho y_0)^2}{2(1-\rho^2)} \right] \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[- \frac{(y_1-\rho y_0)^2}{2(1-\rho^2)} \right] \qquad (4\text{-}14)$$

or

$$= \frac{1}{\sqrt{2\pi}} \exp -(y_1^2/2) \cdot \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(y_0-\rho y_{-1})^2}{2(1-\rho^2)} \right] \frac{1}{\sqrt{2\pi(1-\rho^2)}}$$

$$\exp\left[\frac{(y_{+1}-\rho y_0)^2}{2(1-\rho^2)} \right] \qquad (4\text{-}15)$$

$$= f(y_{-1}) \cdot f(y_o/y_{-1}) \cdot f(y_1/y_0) \qquad (4\text{-}16)$$

Equation 4.16 shows how the Markov sequence emerges. Firstly one ordinate has a value, the second ordinate is conditional on this, and in turn the third ordinate can be considered as being conditional on the second only, and so on.

In the discussion above the profile has been considered as a continuous signal. In the digital presentation of a profile having an exponential autocorrelation function the data becomes a Markov chain and the conditional probability densities become transition probabilities. Thus for a series of ordinates y_{-1}, y_0, y_{+1},.....we are concerned typically, with the probability of the transition from state $i(y_0)$ to state $j(y_{+1})$; this transition probability P_{ij} is an element in the stochastic matrix Π (Papoulis 1966).

$$\Pi = \begin{matrix} P_{11} & P_{12} & \cdots\cdots & P_{1N} \\ \vdots & & & \vdots \\ \cdots & \cdots & \cdots\cdots & P_{NN} \end{matrix} \qquad (4\text{-}17)$$

If the quantisation interval is δy then the transitional probability for the transition from one element of the sequence (having a value between a and (a + δy)) to the subsequent value (having a value between b and (b + δy)) is given by

$$P_{ab} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(b-\rho.a)^2}{2(1-\rho^2)}\right] \delta y \qquad (4\text{-}18)$$

In practice, this approach in discrete rather than continuous terms, can be justified,to some extent, because of the finite quantisation inevitably involved in the statement of ordinate heights in digital analysis. In fact the loss of information involved in current practice is quite small. For example, it has been shown (Widrow 1956) that the loss of information in a signal is quite small even when a coarse quantisation interval is used and this is the basis of the quantisation correlator (Watts 1961).

In general, the Markov property is such that

$$f(y_{+1}y_2 \ldots\ldots y_n) = f(y_{+1}) . f(y_2/y_{+1}) \ldots . f(y_n/y_{n-1}) \qquad (4\text{-}19)$$

or equally with the sequence reversed; the individual elements here are as in equation (4-8).

For the very important case where the members of the sequence, i.e. the ordinates of the waveform, are taken far enough apart to be considered to be independent of each other then equation (4-19) reduces to

$$f(y_{+1}y_2 \ldots . y_n) = f(y) . f(y_2) \ldots . f(y_n) \qquad (4\text{-}20)$$

the number of individual members of the sequence corresponding to the number of degrees of freedom in the waveform, i.e. for a surface having a sharp cut spectrum at B cycles per unit length in the waveform then in a length L of chart the number of degrees of freedom would be L/2B.

The distance corresponding to independence needs some explanation. This does not necessarily mean the same as zero correlation; two points may have zero correlation and yet be related, as for instance two values of a sine wave separated by $\pi/2$. However, for the case where the amplitude distribution is Gaussian then zero correlation does mean independence. Thus, in those cases where the waveform has an autocorrelation function that crosses the zero line (for instance, if the waveform has a spectrum with a short wavelength cut) then the position of zero correlation is unambiguous. However, for cases where the autocorrelation function decays monotonically to zero, for instance the exponential or Gaussian autocorrelation functions, the definition of zero correlation, and consequently of independence, becomes somewhat arbitrary; in this situation one requires a definition of correlation acceptably small such that ordinates having this correlation can be considered as independent. In what follows we shall assume that when the autocorrelation function is exponential and when the correlation has declined to a value of about 0.1 the events can be regarded as independent; this assumption will be discussed and justified later.

The particular value of correlation chosen is not very critical, i.e. the distance corresponding to a value of e^{-1} or

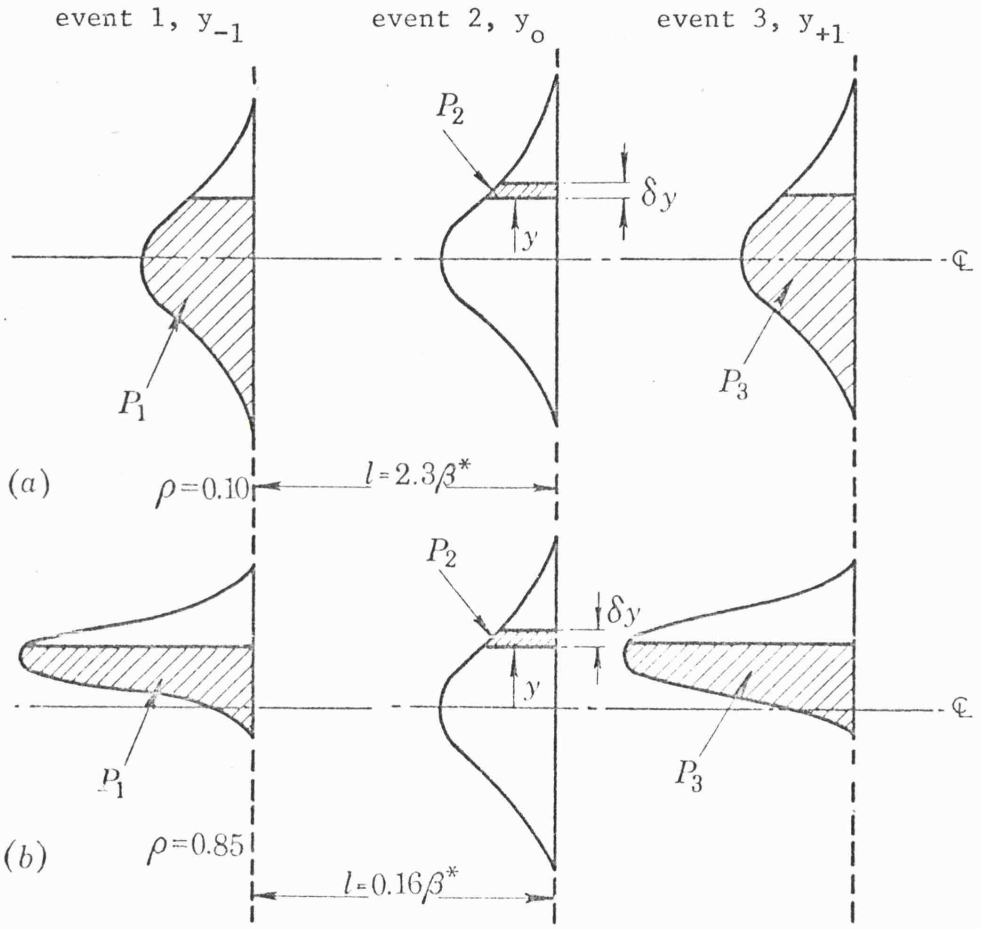
0.5 in addition to 0.1 have been taken by others (Beckmann and Spizzichino 1963, Peklenik and Kubo 1968).

4.3 Digital Analysis

4.3.1 Introduction; three-point analysis

One of the features of the approach outlined above is that the surface is now represented in a form very convenient for the investigation of the results that emerge from the presentation of the outputs from stylus tracer instruments in digital form. The way that this digital information is usually used in investigation of surface waveforms for tribological purposes is by means of three-point analysis. The most widely quoted examples of three-point analysis have used a single sampling interval and this inevitably causes a loss of information compared with that contained in the original waveform. Thus close sampling collects a maximum of information about the profile but the three-point analysis of this data restricts the information to structure on the surface of this same small scale of size. Only by means of more sophisticated techniques such as digital filtering can the total information in the sample be utilised. An alternative is to use differing selections of the same information, by rejecting some data, but still to use three-point e.g. to present information from three-point analysis for a wide range of sampling intervals. However, this can cause problems of rigour. When using the longer sampling intervals one is presenting information about the longer wavelength structure of the waveform obtained by drawing a smooth

Figure 4-5. Model used in deducing distribution of peaks.
(a) Sampling interval, l = 2.3β*; correlation, ρ = 0.10;
(b) sampling interval, l = 0.16β*; correlation, ρ = 0.86.



curve through widely separated points. Therefore any results derived below should bear these reservations in mind.

In the terms of the Markov sequences discussed above this represents investigating the behaviour of three-element Markov sequences. In what follows the derivation of formulae for parameters of fundamental importance in tribology will be considered.

4.3.2 The peak distribution

Consider a sequence of three consecutive ordinates of the profile (figure 4-5(a)). In this diagram and in the discussion which follows the average behaviour of three such consecutive events is considered.

The necessary restrictions on these three events in order to define a peak are as follows:

(a) The central event lies between y and y + δy.

(b) Event 2 has a value of less than y.

(c) Event 3 also has a value of less than y.

We may note that in what follows the valleys can be treated in a similar way. Thus the probability that the central ordinate represents a peak between y and (y + δy) is the multiplication of the probabilities, P_1 P_2 and P_3 where the P's refer to the shaded areas of the height distribution.

Obviously in some situations a peak may require four or even more ordinates adequately to define it but for most purposes these three ordinates will suffice. Also, the assumption is here that the peak defined by these three ordinates has its apex at event 2. The implication of this is discussed in section 4.3.6.

In terms of the joint pr obability density function $f(y_{-1},y_0,y_{+1})$ the condition for a peak as defined above becomes

$$\text{Prob}\left[y_{-1}<y, y<y_0<y+dy, y_{+1}<y\right] = \int_{-\infty}^{y}\int_{y}^{y+\delta y}\int_{-\infty}^{y} f(y_{-1},y_0,y_{+1})\,dy_{-1}dy_0dy_{+1} \tag{4-21}$$

which can be written

$$\int_{y}^{y+\delta y} f(y_0)\int_{-\infty}^{y} f(y_{-1}/y_0)\int_{-\infty}^{y} f(y_{+1}/y_0,y_{-1})\;dy_{+1}\;dy_{-1}\;dy_0 \tag{4-22}$$

and using the Mean Value Theorem

$$= f(y)\int_{-\infty}^{y} f(y_{-1}/y)\int_{-\infty}^{y} f(y_{+1}/y,y_{-1})\;dy_{+1}\;dy_{-1}\;dy \tag{4-23}$$

Which is the general equation for a peak as defined by three-point analysis.

For an exponential correlation function equation (4-23) reduces to

$$f(y)\int_{-\infty}^{y} f(y_{-1}/y)\;dy_{-1}\int_{-\infty}^{y} f(y_{+1}/y)\;dy_{+1}\;dy \tag{4-24}$$

Hence the probability density* that an ordinate is a peak at height y is given by

$$f^* \ (y,\rho) = f(y) \int_{-\infty}^{y} f(y_{-1}/y) \ dy_{-1} \int_{-\infty}^{y} f(y_{+1}/y) \ dy_{+1} \qquad (4\text{-}25)$$

Inserting (4-8) into this yields

$$f^* \ (y,\rho) = \frac{\exp \ (-y^2/2)}{\sqrt{2\pi}} \ \frac{1}{2\pi(1-\rho^2)} \int_{\infty}^{y}$$

$$\exp \left[- \frac{(y_{-1}-\rho y \ )^2}{2(1-\rho^2)} \right] dy_{-1} \int_{-\infty}^{y} \exp \left[- \frac{(y_{+1}-\rho y \ )^2}{2(1-\rho^2)} \right] dy_{+1}$$

$$(4\text{-}26)$$

Thus

$$f^* \ (y,\rho) = \frac{1}{4\sqrt{2\pi}} \left[1 + \text{erf} \ (y/\sqrt{2} \sqrt{\tfrac{1-\rho}{1+\rho}}) \right]^2 \exp \ (-y^2/2)$$

$$= \frac{1}{\sqrt{2\pi}} \ \Phi^2 \ (y \sqrt{\tfrac{1-\rho}{1+\rho}}) \ \exp \ (-y^2/2)$$

$$(4\text{-}27)$$

When the ordinates are spaced so far apart that they can be regarded as independent of each other equation (4-27) reduces to

$$f^* \ (y) = \frac{1}{4\sqrt{2\pi}} \left[1 + \text{erf} \ (y/\sqrt{2}) \right]^2 \exp \ (-y^2/2)$$

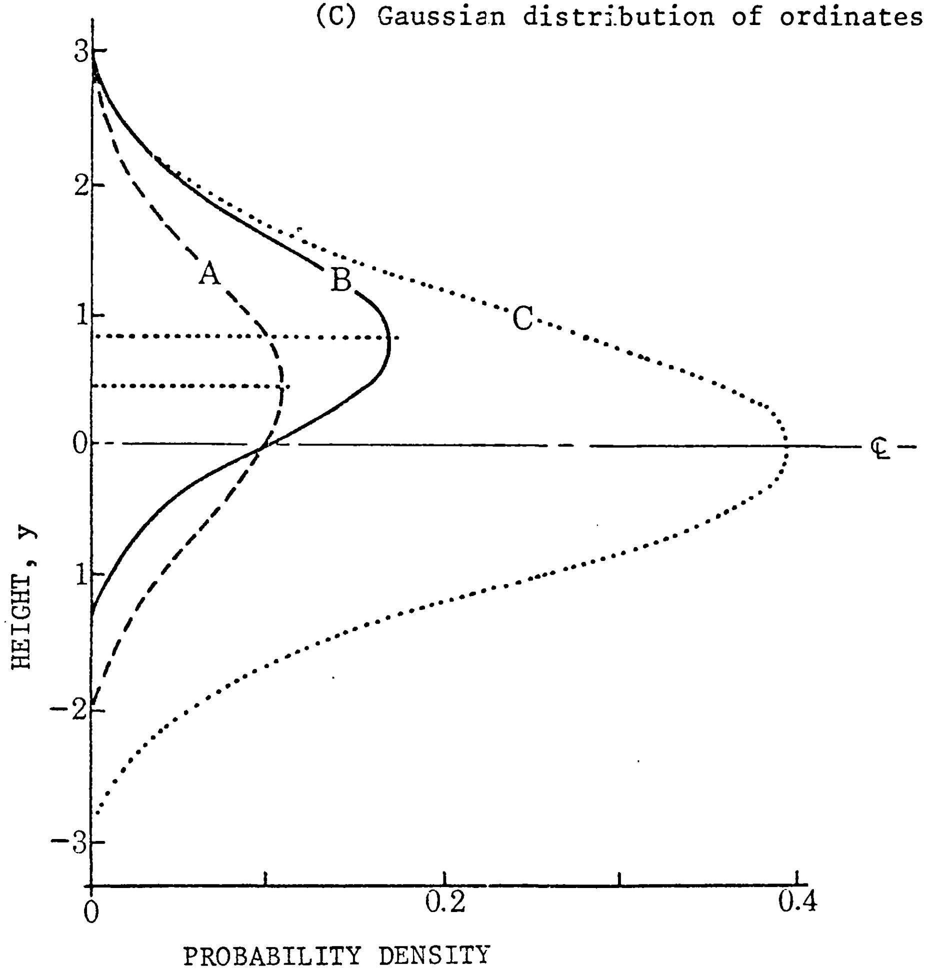
$$= \frac{1}{\sqrt{2\pi}} \ \Phi^2(y) \ \exp \ (y^2/2)$$

$$(4\text{-}28)$$

* Strictly in what follows the expressions are only densities when multiplied by a constant factor to make the integral unity. However, this is taken into account when moments are taken.

Figure 4-6. Probability densities of an ordinate being a peak at height y. The height, y, is normalised by the RMS valie (σ) of the ordinates. Results are shown for two different values of the sampling interval (l).

(A) $l = 2.3\beta^*$; correlation, $\rho = 0.10$; average peak height = 0.82;
(B) $l = 0.16\beta^*$; correlation, $\rho = 0.86$; average peak height = 0.41;
(C) Gaussian distribution of ordinates.



It should be noted that this equation is not dependent upon the assumption of an exponential correlation function.

The general expression for a peak using three-point analysis for an arbitrary correlation function is obtained by putting equations (4-10) into (4-22) and using the identity

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(t-m)^2}{2\sigma^2}\right] dt = \frac{1}{2}\left[1 + \operatorname{erf}\frac{(x-m)}{\sigma\sqrt{2}}\right]$$

thus

$$f^*\,(y,\rho_1\rho_2) = \frac{\exp\,(-y^2/2)}{\sqrt{2\pi}} \; \frac{1}{\sqrt{2\pi(1-\rho_1{}^2)}} \int_{-\infty}^{y}$$

$$\exp\left[-\frac{(y_1-\rho_1 y)^2}{2(1-\rho_1{}^2)}\right] \times \frac{1}{2}\left[1 + \operatorname{erf}\left[\frac{y(1-\rho_1+\rho_1\rho_2)-y_{-1}(\rho_1{}^2-\rho_2)}{\dfrac{2(1-\rho_2)(1+\rho_2-2\rho_1{}^2)}{(1-\rho_1{}^2)}}\right]\right] dy_{-1}$$

(4-29)

Figure 4-5(a) shows the three-point model for an exponential correlation function for spacings which are large and consequently the ordinates are independent; figure 4-5(b) shows the corresponding situation when the events are closer and must influence each other. Figure 4-6 shows plots of the derived forms of the peak height distribution for a high and low value of correlation ρ; the height distribution of the ordinates is also shown for comparison. The trends with varying values of ρ will be observed. As $\rho\to 0$ (large sampling interval) the shape of the peak height distribution becomes slightly skewed, its mean value approaches 0.85 and its

standard deviation approaches a value of 0.7. Thus, when using larger sampling intervals the main, longer wavelength structure of the profile is revealed (neglecting here the problem of aliasing), and the peaks tend to be above the centre line.
As ρ→1 the shape of the peak height distribution and its mean value and standard deviation, approaches those of the height distribution of the ordinates. Thus, when using short sampling intervals one is concerned with the shorter wavelength structure of the profile.

The mean value of the peak height density curve $\overline{y}^*$ (ρ) is found by taking the first moment of f* (y,ρ) in the normalised version of equation (4-27). Thus

$$\overline{y}^*(\rho) = \frac{\int_{-\infty}^{\infty} \frac{\exp(-y^2/2)}{\sqrt{2\pi}} \frac{y}{4} \cdot \left[1 + \text{erf}\ (y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right]^2 dy}{\int_{-\infty}^{\infty} \frac{\exp(-y^2/2}{\sqrt{2\pi}} \frac{1}{4} \left[1 + \text{erf}\ (y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right]^2 dy} \qquad (4\text{-}30)$$

which gives

$$\overline{y}^*\ (\rho) = \frac{1}{2N} \left(\frac{1-\rho}{\pi}\right)^{1/2} \qquad (4\text{-}31)$$

where N is the ratio of the number of peaks to ordinates and is given by

$$N = \frac{1}{\pi} \tan^{-1} \sqrt{(3-\rho)/(1+\rho)} \qquad (4\text{-}32)$$

Similarly the variance $(\sigma)^2$ of the peak heights is the second central moment. Thus

$$\left[\sigma^*(\rho)\right]^2 = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{\exp(-y^2/2)}{\sqrt{2\pi}} \frac{(y-\bar{y}^*(\rho))^2}{4}\left[1+\operatorname{erf}\,(y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right]^2 dy}{\displaystyle\int_{-\infty}^{\infty} \frac{\exp(-y^2/2)}{\sqrt{2\pi}} \frac{1}{4}\left[1+\operatorname{erf}\,(y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right]^2 dy} \tag{4-33}$$

$$\left[\sigma^*(\rho)\right]^2 = \left[\frac{\frac{1}{\pi}\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}} + \frac{1}{2\pi}(1-\rho)\sqrt{\frac{1+\rho}{3-\rho}}}{\frac{1}{\pi}\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}} - \left[\frac{\frac{\sqrt{1-\rho}}{2\sqrt{\pi}}}{\frac{1}{\pi}\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}}\right]^2\right] \tag{4-34}$$

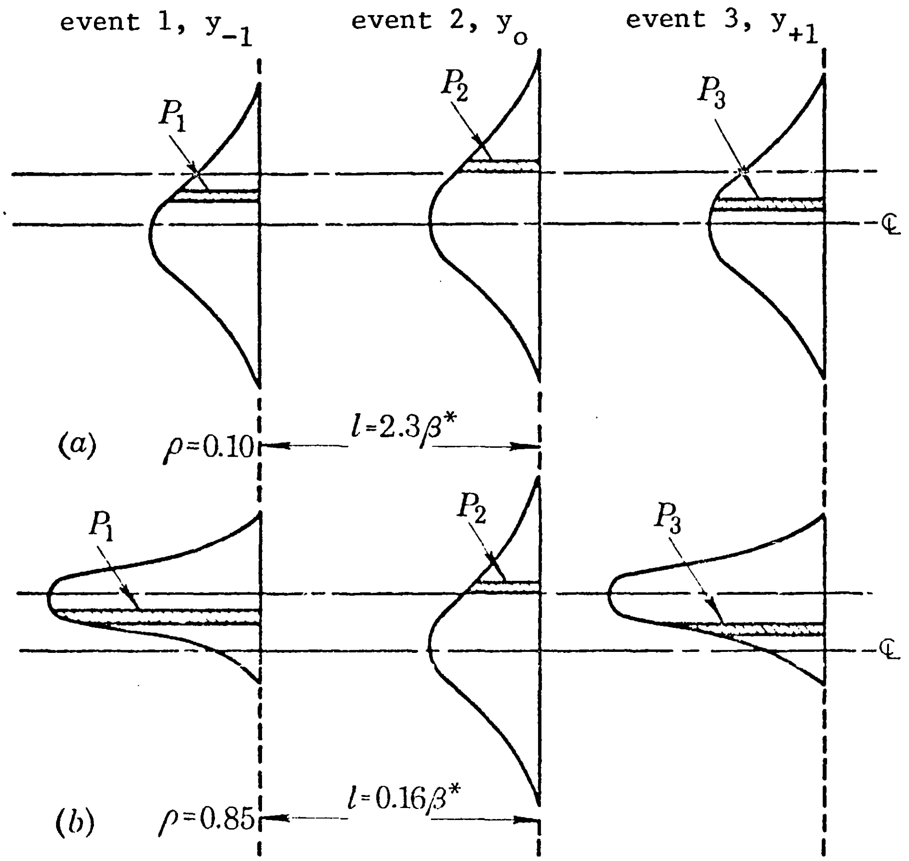
$$\therefore\ \sigma^*(\rho) = \sqrt{1 + \frac{(1-\rho)\sqrt{1+\rho}}{2\sqrt{3-\rho}\,\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}} - \frac{\pi(1-\rho)}{4\left[\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}\right]^2}} \tag{4-35}$$

$$\sigma^*\,(\rho)\quad \sqrt{1 + \frac{6}{\pi}\ \frac{(1-\rho)\sqrt{1+\rho}}{(4-\rho)\sqrt{3-\rho}} - \frac{36\,(1-\rho)}{\pi(4-\rho)^2}} \tag{4-37}$$

It should be noted that equations (4-27) and (4-28) when divided by equation (4-32) are true probability density functions of peak heights.

Equation (4-32) shows that as the correlation ρ increases from zero to unity N falls from 1/3 to 1/4. These limiting values have a simple explanation. As the sampling interval is increased

Figure 4-7. Model used in deducing the distribution of curvatures.
(a) Sampling interval, $l = 2.3\beta^*$; correlation, $\rho = 0.10$;
(b) sampling interval, $l = 0.16\beta^*$; correlation, $\rho = 0.86$.



$\rho \to 0$ and $N \to 1/3$, the three events are then effectively independent (figure 4-5(a)) and the chance that any one of them (e.g. the centre one) is the highest one becomes one third. On the other hand as the sampling interval is decreased $\rho \to 1$ and $N \to 1/4$. The modified distributions of the two outer events are now centred on the central ordinate (figure 4-5(b)); the areas P_1 and P_3 have values of 1/2 and the probability that the central event in a peak is 1/4.

4.3.3 Peak curvature; some general assumptions

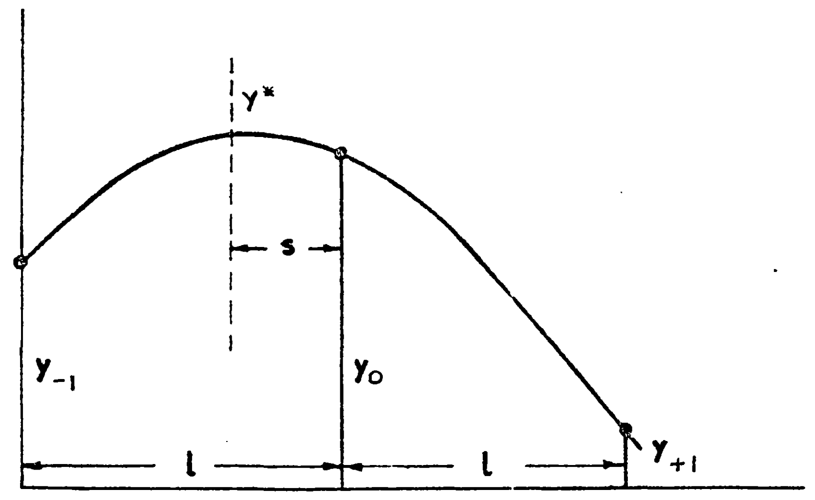
To provide an adequate description of a surface in terms of a distribution of asperities it is also necessary to specify their radii of curvature. It is more convenient to discuss this in terms of a distribution of curvatures and the method adopted by Greenwood and Williamson in deriving curvatures from the digital information will be adhered to. The assumption here is that one is justified in fitting a parabola to the profile by three-point analysis.

The problems involved in this assumption and the limits within which this analysis are justified are complex. They will be briefly discussed here. However, it may be noted that the results obtained are subject to the limitations discussed below.

Figure 4-7(a) shows one possible arrangement of three events which will give a peak at height y with a curvature C given by

FIG 4-8

DEFINITION OF CURVATURE USING A PARABOLIC FIT.



$$C = 2y_0 - y_{+1} - y_{-1} \qquad (4\text{-}38)$$

A negative sign being here omitted to correspond to a convention that a peak has a positive curvature.

A number of assumptions are made in arriving at equation (4-38), they will be taken one by one. Firstly, the curvature is assumed to be given by the second differential. This, in general, is a valid argument because

$$\text{curvature} = \frac{\frac{d^2y}{dx^2}}{\left[1 + (\frac{dy}{dx})^2\right]^{3/2}} \qquad (4\text{-}39)$$

which reduces to curvature $= \frac{d^2y}{dx^2}$ when $\frac{dy}{dx}$ is small, a situation likely to be true in regions near to the peaks.

The next assumption is that y_0 is always at the apex of a vertical parabola, i.e. the highest point of the peak is at y_0. In general, as figure 4-8 shows, the apex y* of a vertical parabola is shifted relative to the ordinate of y_0 by the shift s

$$\text{where } s = \frac{\ell/2\ (y_{+1}-y_{-1})}{2y_0-y_{+1}-y_{-1}} \qquad (4\text{-}40)$$

In addition, the height of the peak, hitherto assumed to be y_0, is at the apex of the parabola at y*. Thus the error in the original assumption is given by

$$y^*-y_0 = \frac{1}{8}\frac{(y_{+1}-y_{-1})^2}{(2y_0-y_{-1}-y_{+1})} \qquad (4\text{-}41)$$

For a typical case if y_0 is 3, y_1 is + 1 and y_{-1} is -1 then the error in the height is + $\frac{1}{12}$ which taken as a percentage of the normal range of the surface (assumed to be 6) works out at less than a two percent error. (All heights normalised by the RMS value σ). For a given degree of asymmetry between y_{+1} and y_{-1} the error gets progressively smaller as y_0 gets higher, a fact of considerable practical importance because it is at the higher peaks that the contact occurs. However, in all this the basic assumption is that an estimate of the curvature, as usually defined, can be made using equation (4-38), an assumption which will be progressively less severe as $\rho \to 1$, when the fourth order central differences become small i.e. the curvature at y_0

$$= -(2y_0-(y_{-1} + y_{+1})) - \frac{1}{12} \delta^4 y_0 \qquad (4\text{-}42)$$

Given the assumptions which have been discussed above, then the probability of the configuration for the closely correlated and independent cases is given by P_1 x P_2 x P_3 where P_1, P_2 and P_3 are the areas shown in figures 4-7(a) and (b)

4.3.4 Peak Curvature

In general the probability of an ordinate being a peak at height between y and y + δy and of curvature C covering all possible configurations will be given by

$$f^*(y,C,\rho) = \int_{y}^{y+\delta y} \int_{y-C}^{y} f(y_0),f(y_{-1}/y_0) \, . \, f(y_{+1}=2y_0-y_{-1}-C/y_0,y_{-1}) \, dy_{-1} \, dy_0$$

(4-43)

from which the probability density of an ordinate being a peak at height y with curvature C is

$$f^* \ (y,C,\rho) = f(y) \ . \int_{y-C}^{y} f(y_{-1}/y) \ . \ f(\frac{2y-y_{-1}-C}{y, \ y_{-1}}) \ dy_{-1} \qquad (4\text{-}44)$$

This is an expression giving the curvature with three-point analysis for a general autocorrelation function.

For the exponential autocorrelation function, equation (4-44) becomes

$$f^* \ (y,C,\rho) = \frac{1}{\sqrt{2\pi}} \exp\ (-y^2/2) \int_{y-C}^{y} \frac{1}{2\pi(1-\rho^2)} \exp\left[-\frac{(y_{-1}-\rho y)^2}{2(1-\rho^2)}\right] \cdot \exp\left[-\frac{(2y-y_{-1}-C-\rho y)^2}{2(1-\rho^2)}\right] dy_{-1}$$

(4-45)

which is a convolution integral thus enabling simple graphical equivalents to be drawn. Equation (4-45) reduces to:

$$f^*(y,C,\rho) = \frac{\exp(-y^2/2)}{2\pi\sqrt{2(1-\rho^2)}} \exp\left[-\frac{|(1-\rho)y-C/2\ |^2}{(1 - \rho^2)}\right] \text{erf} \left[\frac{C}{2(1-\rho^2)}\right]$$

(4-46)

which for $\rho = 0$ becomes

$$f^* \ (y,C) = \frac{\exp(-y^2/2)}{2\pi\sqrt{2}} \exp \left[-(y-C/2\ )^2\right] \text{erf}\ (C/2) \qquad (4\text{-}47)$$

The probability density function that any ordinate is a peak of curvature C at any height is obtained direct from equation (4-46) by integrating, giving

$$f^*(C,\rho) = \left[\frac{1}{4\pi(3-\rho)(1-\rho)}\right]^{1/2} \exp\left[\frac{-C^2}{4(3-\rho)(1-\rho)}\right] \operatorname{erf}\left[\frac{C}{2\sqrt{1-\rho 2}}\right] \quad (4\text{-}48)$$

and for $\rho = 0$

$$f^*\,(C) = \frac{1}{\sqrt{12\pi}} \exp\left(\frac{-C^2}{12}\right) \operatorname{erf}\,(C/2) \quad (4\text{-}49)$$

These distributions of equation (4-48) and(4-49) are skewed towards zero curvature as found experimentally by Greenwood and Williamson (1966) from digital analysis of surfaces generated by random processes such as bead blasting. These authors suggest a Gamma function as a suitable description of the distribution but these equations are nearer to a Rayleigh distribution and for large curvatures become very nearly Gaussian. A comparison of these equations with results derived from surface profiles will be given in Section 4.3.7. In Section 4.3.6 the relationship of equation (4-49) to a Rayleigh distribution is discussed again.

The mean curvature $\overline{C}^*$ for all peaks is obtained by finding the first moment of f* (C,ρ) in equation (4-48).
Thus

$$\overline{C}^*\,(\rho) = (3-\rho)\,\sqrt{1-\rho}\Big/ 2N\,\sqrt{\pi} \quad (4\text{-}50)$$

where the expression has been normalised by N the ratio of peaks to ordinates (equation 4-32).

4.3.5 Other parameters

The curvature (or strictly) the second differential of the profile as a whole is given by

$$f(C,\rho) = \frac{1}{\sqrt{4\pi(3-\rho)(1-\rho)}} \exp\left[\frac{-C^2}{4(3-\rho)(1-\rho)}\right] \qquad (4-51)$$

Equation (4-51) is a Gaussian distribution having a mean of zero and a standard deviation of $\sqrt{2(3-\rho)(1-\rho)}$. A simple check on this value of the standard deviation is obtained by finding the variance of curvature from equation (4-35).

Thus

$$E\left[2y_0-(y_{-1}+y_{+1})\right]^2 = 6 - 8\rho + 2\rho^2 \qquad (4-52)$$

The distribution of slopes as well as curvature is important because one widely used criterion for the onset of plastic flow (Blok 1952, Halliday 1955) uses the mean slope of the flanks of the asperities. The slope distribution of the profile $f(m,\rho)$is easily obtained as in the case of curvature of the profile as a whole because the formula itself involves a simple linear relationship of the Gaussian variates y_{-1} and y_{+1} and so is itself Gaussian with mean zero and variance given by $2(1-\rho^2)$.

Hence

$$f(m,\rho) = \exp\left[\frac{-m^2}{4(1-\rho^2)}\right] \Big/ \left[4\pi(1-\rho^2)\right]^{1/2} \tag{4-53}$$

from which the average modulus of the slope $\overline{m}$ can be obtained taking into account the spacing and variance to give

$$\overline{m} = \frac{\sigma}{\ell} \cdot \left[\frac{1-\rho^2}{\pi}\right]^{1/2} \tag{4-54}$$

4.3.6 Effect of assumptions on curvatures*

In this section the effect of the assumption that the apex of the peak occurs at y_0 will again be considered. From figure 8, if R is the radius of curvature

$$y_{-1} = y^* - \frac{1}{2R}(\ell+s)^2$$

$$y_0 = y^* - \frac{1}{2R}s^2$$

$$y_{+1} = y^* - \frac{1}{2R}(\ell-s)^2$$

using the spherometer formula approximation to a true circular peak.

* I am indebted to Mr. A. Dyson (Shell Research Ltd.), Adrian Visiting Fellow of the University, for suggesting the approach used in this section.

Letting $\frac{1}{R} = \frac{C}{\ell^2}$ where $C = 2y_0 - y_1 - y_{-1}$

$$y_{-1} = y^* - \frac{C}{2\ell^2}(\ell+s)^2$$

$$y_0 = y^* - \frac{C}{2\ell^2} s^2$$

$$y_{+1} = y^* - \frac{C}{2\ell^2}(\ell-s)^2$$

Now the probability density of a peak of curvature C at a true maximum height of y* whose apex is a distance s from y_2 is given by $\overset{*}{f}$ (y*,C,s). This can be determined from the co-ordinate space y_{-1}, y_0, y_{+1} by a transformation whose Jacobian J is given by $\frac{\partial(y_{-1}, y_0, y_{+1})}{\partial(y^*, C, s)}$

$$\text{where } J = \begin{vmatrix} \frac{\partial y_{-1}}{\partial y^*}, & \frac{\partial y_{-1}}{\partial C}, & \frac{\partial y_{-1}}{\partial s} \\ \frac{\partial y_0}{\partial y^*}, & \frac{\partial y_0}{\partial C}, & \frac{\partial y_0}{\partial s} \\ \frac{\partial y_{+1}}{\partial y^*}, & \frac{\partial y_{+1}}{\partial C}, & \frac{\partial y_{+1}}{\partial s} \end{vmatrix} \qquad (4\text{-}56)$$

$$J = \begin{vmatrix} 1, & -\frac{1}{2\ell^2}(\ell+s)^2, & -\frac{C}{\ell^2}(\ell+s) \\ 1, & -\frac{1}{2\ell^2}s^2, & -\frac{C}{\ell^2}s \\ 1, & -\frac{1}{2\ell^2}(\ell-s)^2, & +\frac{C}{\ell^2}(\ell-s) \end{vmatrix}$$

$$= C/\ell \qquad (4\text{-}57)$$

If we now consider the case where y_{-1}, y_0 and y_{+1} are all independent then their joint probability density is $f(y_{-1},y_0,y_1)$ given by

$$\frac{1}{(2\pi)^{3/2}} \exp\left[-\frac{y_{-1}^2 + y_0^2 + y_{+1}^2}{2}\right]$$

Using the transformation

$$f(y_{-1},y_0,y_1) \cdot J = \overset{*}{f}(y^*,C,s) \qquad (4\text{-}58)$$

(for a justification see Davenport and Root 1958) yields

$$\overset{*}{f}(y^*,C,s) = \frac{C}{\ell(2\pi)^{3/2}}$$

$$\exp\left[-\left((y^*-\frac{C}{2\ell^2}(\ell+s)^2)^2 + (y^*-\frac{C}{2\ell^2}s^2)^2 + (y^*-\frac{C}{2\ell^2}(\ell-s)^2)\right)\right]$$

$$= \frac{C}{\ell(2\pi)^{3/2}} \exp\left[-\frac{(3y^{*2}-\frac{C}{\ell^2}y^*(2\ell^2+3s^2) + \frac{C^2}{\ell^4}(2\ell^4+3s^4+12\ell^2s^2)}{2}\right]$$

$$(4\text{-}59)$$

Now the total curvature distribution of the peaks is obtained by integrating with respect to y^* to give $\overset{*}{f}(C,s)$. Thus

$$\overset{*}{f}(C,s) = \frac{C}{\ell 2\pi\sqrt{3}} \exp\left[-\frac{C^2}{12\ell^4}(\ell^4+12\ell^2s^2)\right] \qquad (4\text{-}60)$$

which is a Rayleigh distribution - demonstrating the existence of a Rayleigh distribution of curvatures rather than a Gamma function as suggested as a possible distribution by Greenwood and Williamson (see discussion of equation (4-49) above). To take into account all the possible peak apex positions this equation has to be integrated with respect to s. The limits of integration corresponding to s = $\pm\ell/2$ have to be used because within these limits the curvature is positive if $y_0 > y_{+1}, y_{-1}$ which corresponds to our definition of a peak. (see equations (4-55)). Then,

$$\overset{*}{f}(C,s) = \frac{C}{\ell 2\pi\sqrt{3}} \int_{-\ell/2}^{\ell/2} \exp\left(-\frac{C^2}{12}\right) . \exp\left(-\frac{C^2 s^2}{\ell^2}\right) ds$$

$$= \frac{1}{\sqrt{12\pi}} \exp\left(-\frac{C^2}{12}\right) . \operatorname{erf}(C/2) \qquad (4\text{-}61)$$

As can be seen equation (4-61) is identical with equation (4-49) illustrating that as far as the curvature of all the peaks is concerned the errors due to this particular assumption are zero. Exactly the same argument can be followed in the correlated case.

4.3.7 Analysis of surface profiles

The validity of the theory given above has been checked using digital analysis of profile meter outputs. For this purpose Aachen 64-13, a typical ground surface used in an O.E.C.D. research programme was used. The experimental results were derived from surface profiles having been obtained by data logging the output from both a Talysurf 4 and a Talystep as explained in Section 3. The results presented are based upon five profiles each

TABLE 4.1

Relation between sampling interval (ℓ) and correlation (ρ) between successive samples on Aachen 64-13.

Sampling interval (μm)	15	6.0	3.0	2.0	1.0	0.5	0.25
Correlation (ρ)	0.10	0.40	0.63	0.74	0.86	0.92	0.96

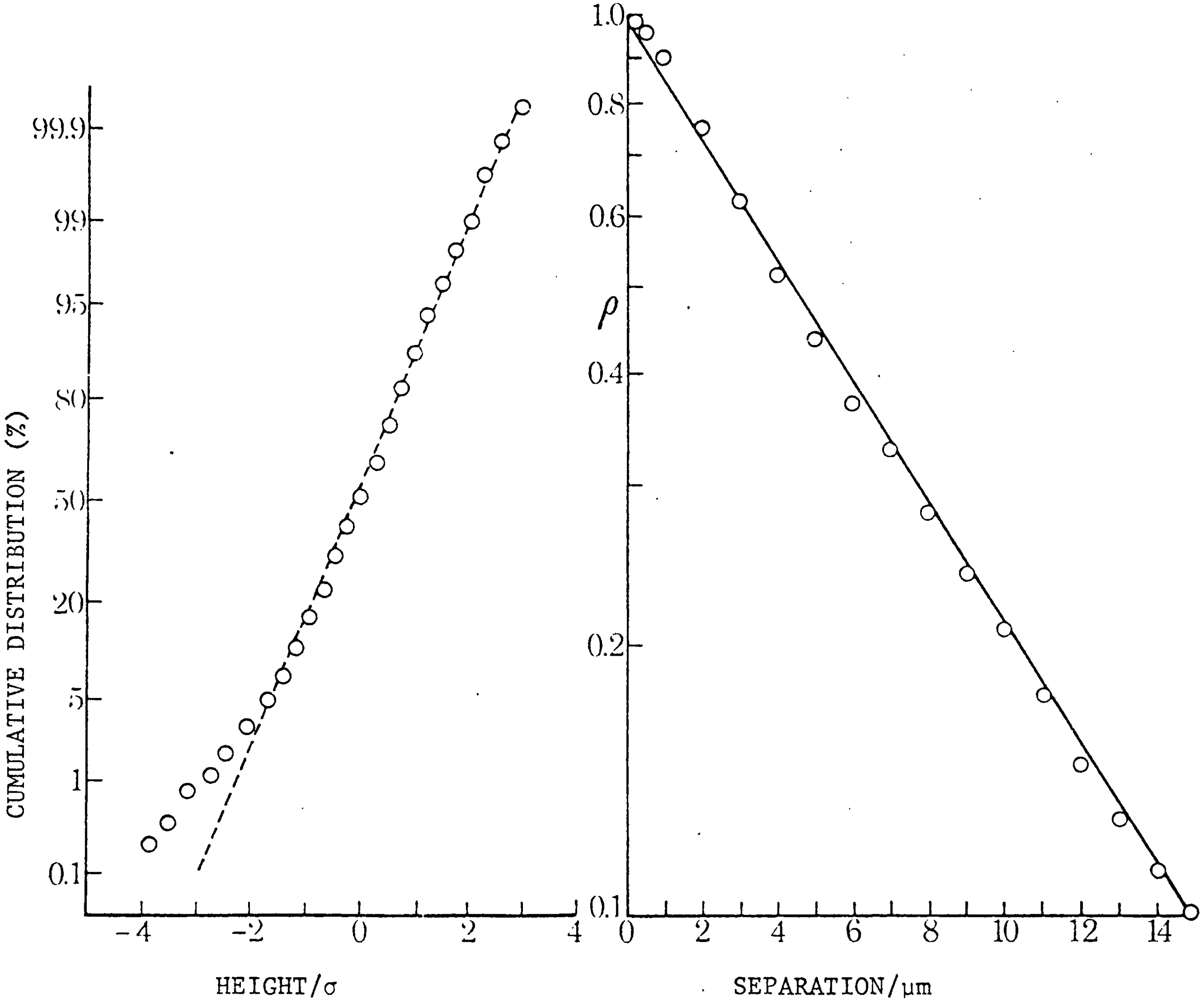
TABLE 4.2

Showing second order joint distribution practically for Aachen 64-13. Joint distribution of Aachen 64-13 between adjacent ordinates.

Sampling interval (μm)	1.25	2.5	15
Correlation ρ	0.825	0.68	0.1
σ practical	0.96	0.67	0.52
σ theoretical	0.99	0.73	0.565

The results in the Table were obtained from 1000 data points. In the particular record of Aachen 64-13 used in this exercise the sampling rates were in rational units of micro-inches rather than micrometres; the 2.5 μm and 1.25 μm corresponding to 100 micro-inch and 50 micro-inch spacing respectively.

Figure 4-9. Characteristics of profiles of Aachen 64-13 represented as a random signal. (a) Cumulative distribution of heights (normal probability paper); (b) Correlation as a function of sampling interval (logarithmic-linear plot).



consisting of some 10,000 ordinates. Statistical analysis shows that the normalised standard errors of the results presented below are approximately 2% for mean values, and 5% for individual points on the probability distributions.

By suitable selection of data in the computer it was possible to present results for sampling intervals between 0.25 μm and 15 μm. It was thus found (figure 4-9) that the model used was a good representation of the data obtained from the surface profiles; the distribution of ordinates was close to Gaussian with an RMS value (σ) of 0.5 μm and the autocorrelation function was close to exponential with a correlation distance (β*) of 6.5 μm. In subsequent discussion the theoretical values of the correlation (ρ) are used for the selected values of the sampling interval as shown in Table 4.1. It will be observed that any divergences between these values and those obtained from the profiles are, for the most part, within the limits of experimental error. Figure 4-10 shows the joint distribution of two ordinates on Aachen 64-13 obtained from the analysis of this same data and Table 4.2 provides a comparison of theory and experiment for this material.

In the results presented in graphical form, sampling intervals (ℓ) of 15, 6, 3, 2 and 1 μm (corresponding to correlations of 0.10, 0.40, 0.63, 0.74 and 0.86 respectively) have been selected to display certain important features. Figure 4-11 shows a comparison of theory and experiment for the probability that an ordinate is a peak at height y (equation (4-27). It will be observed that for

FIG 4-10

CONDITIONAL DENSITY FUNCTIONS AACHEN 64-13 (1000 ORDINATES)

(c) EVERY ORDINATE (b) EVERY OTHER ORDINATE (a) EVERY TWELFTH ORDINATE.

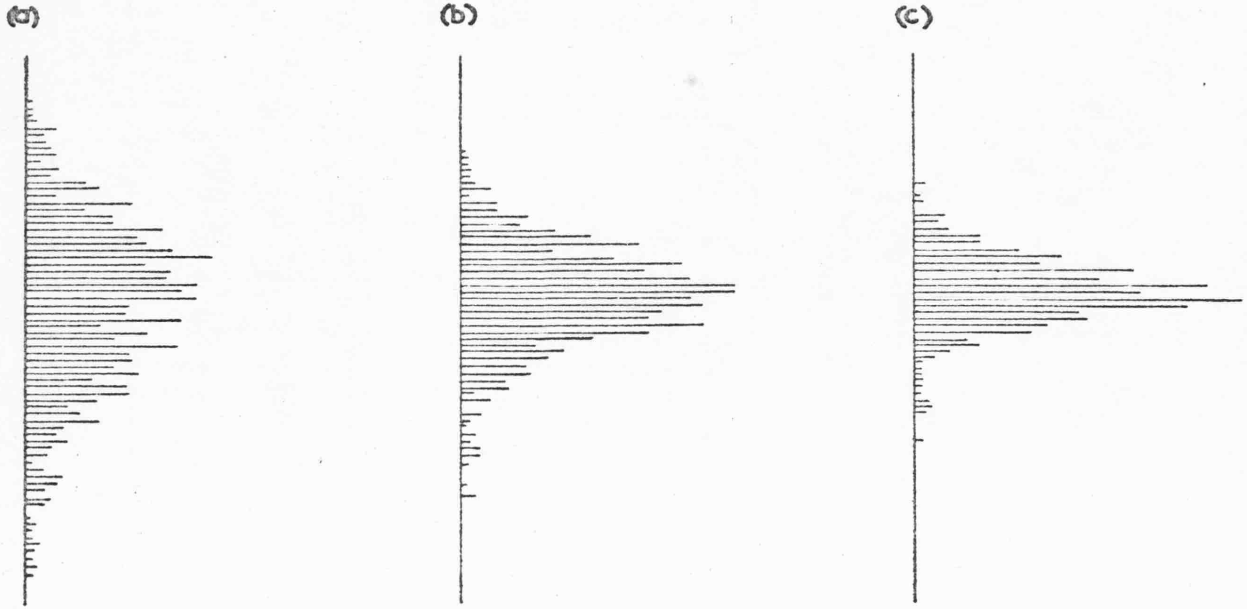


Figure 4-11(a). Peak distribution correlation 0.10.

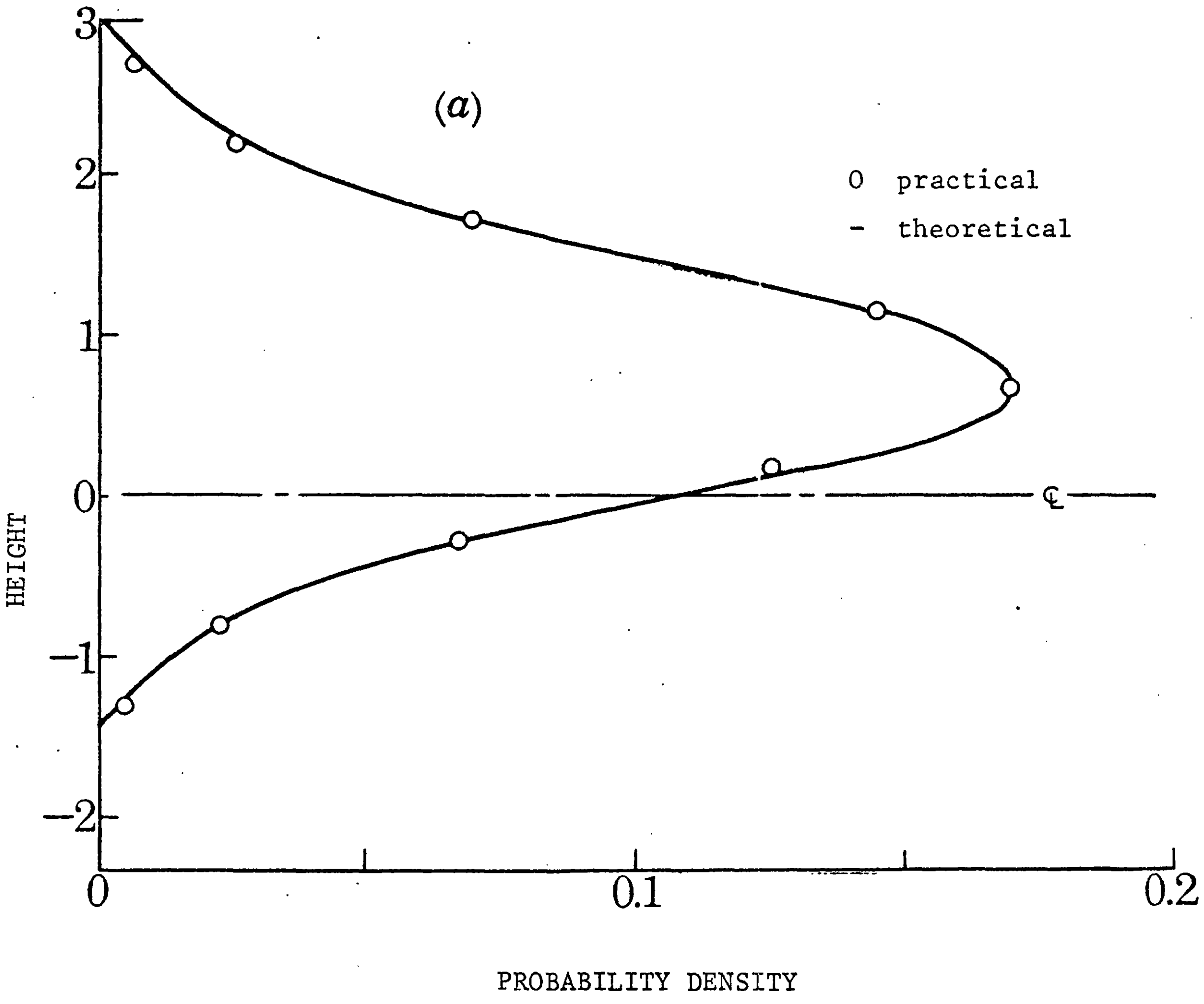


FIG. 4 -11 b

PEAK DISTRIBUTION CORRELATION 0·4

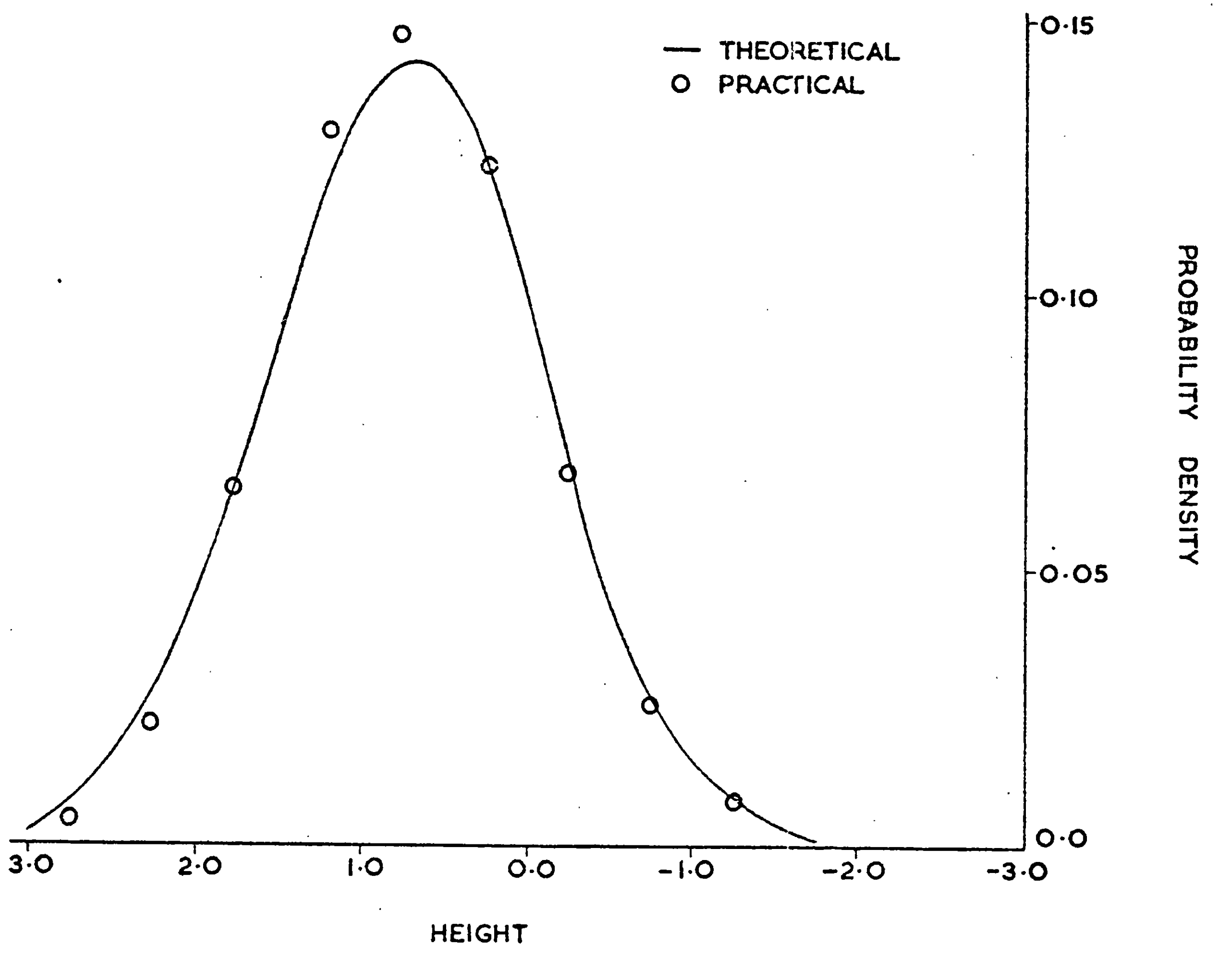


Figure 4-11(c). Peak distribution correlation 0.63

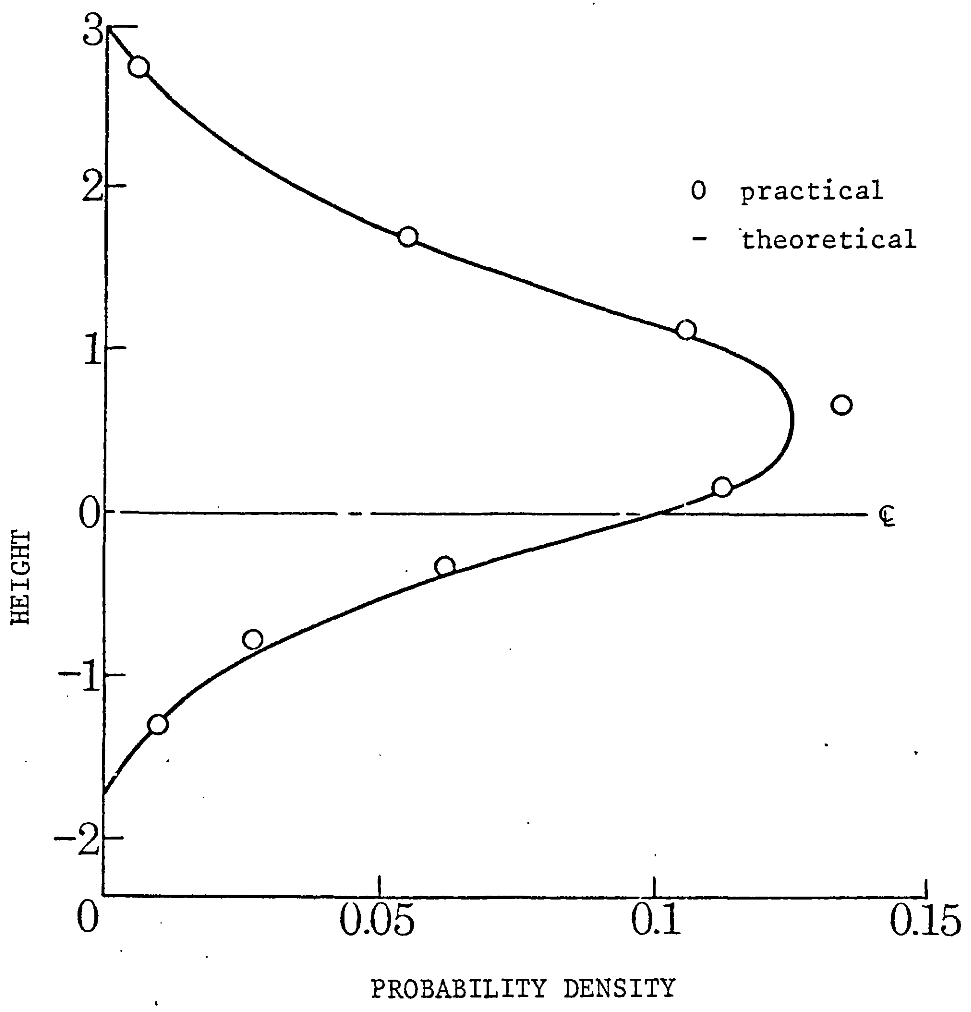


FIG 4-11d

PEAK DISTRIBUTION CORRELATION 0·74.

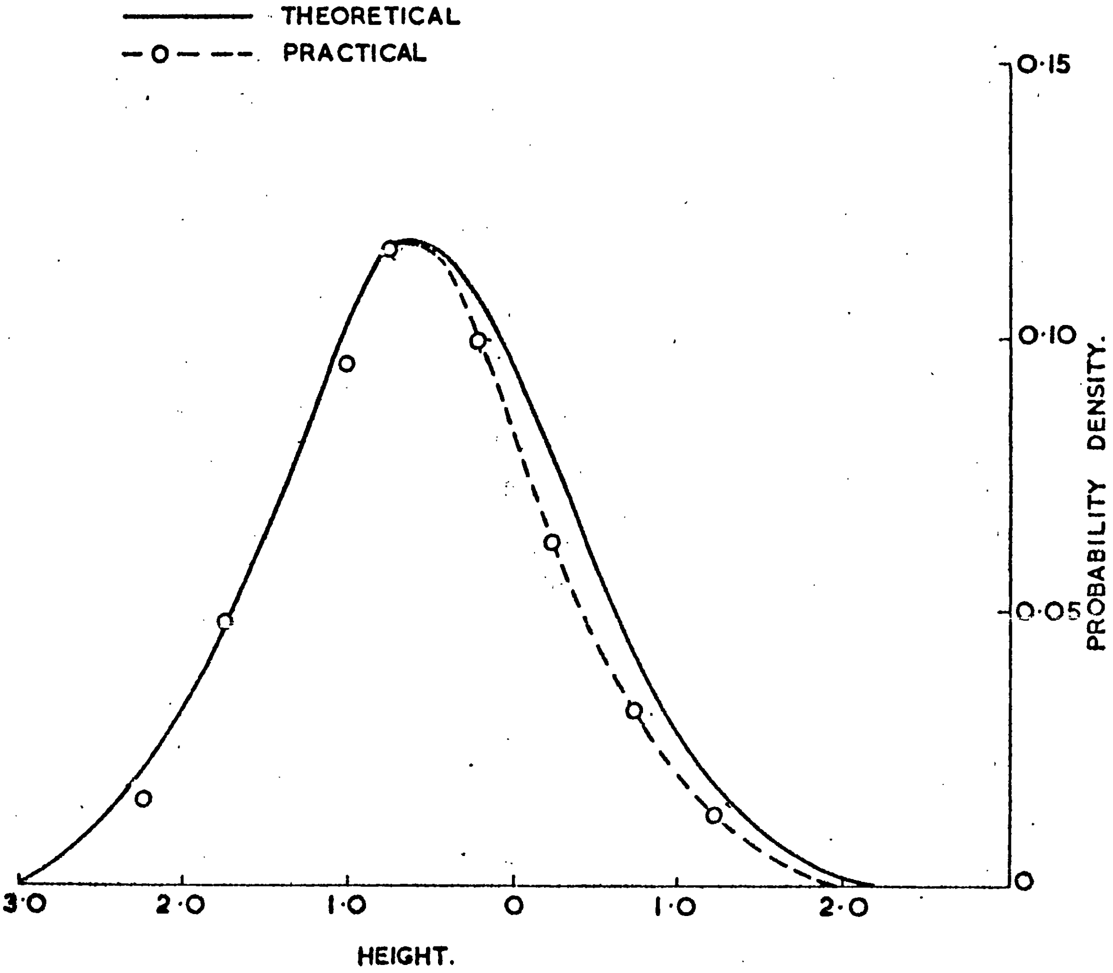


Figure 4-11(c). Peak distribution correlation 0.86

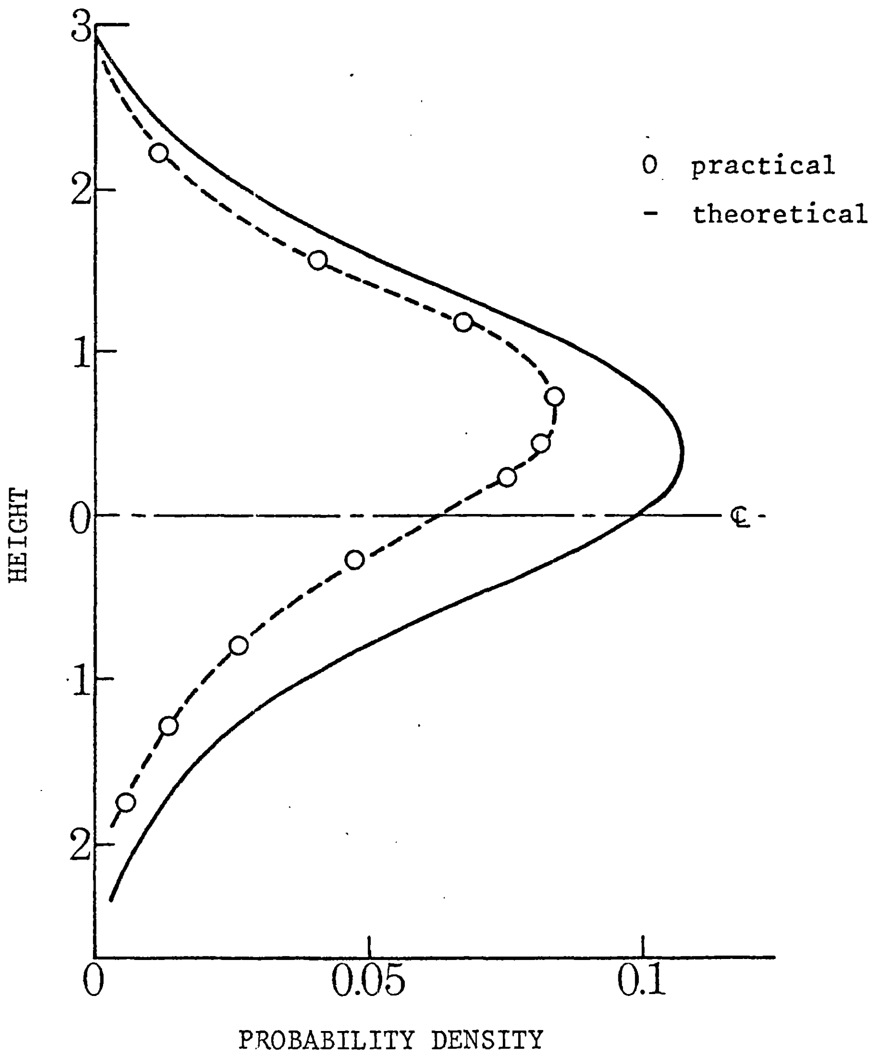
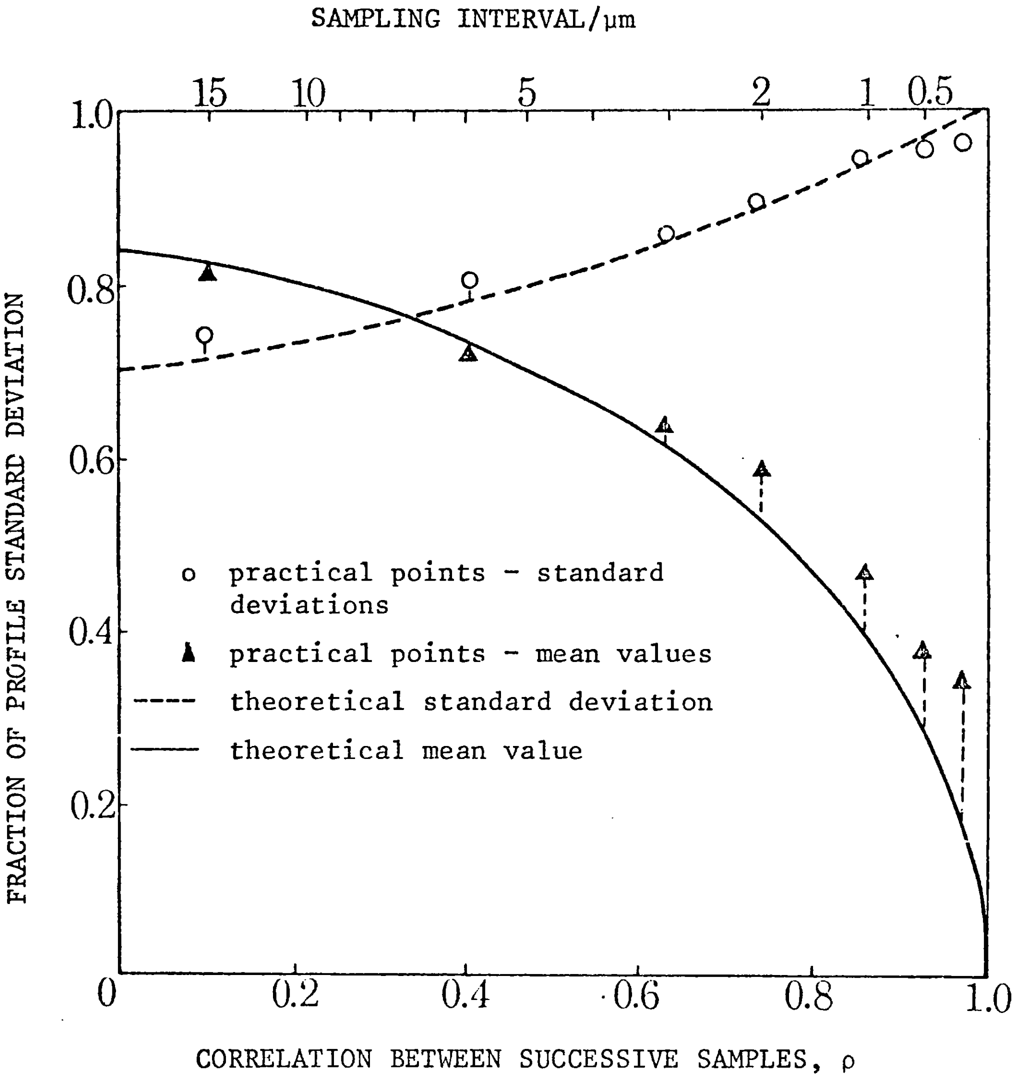


Figure 4-12. Characteristics of the distribution of peaks. The full line gives the mean value and the broken line the standard deviation; they are normalised by the standard deviation of the ordinates (σ). The experimental points are derived from digital analysis of profiles from Aachen 64-13.



ℓ = 15 μm down to ℓ = 3μm the agreement between theory and experiment is good. However, for ℓ = 1 μm, figure 4-11(e), there is a marked divergence, the number of peaks detected falling significantly below the theoretical values. The results for all values of the sampling interval are shown in figure 4-12 in which the mean value and the standard deviation of the distribution of peaks, equations (4-31) and (4-35), are plotted against the value of the correlation between successive samples. The most significant divergence between theory and experiment is the fact that, for the shorter sampling intervals the mean values lie above the theoretical predictions (see also figure 4-11(e)).

Figure 4-13 presents theory and experiment for the probability that an ordinate is a peak of given curvature. As before, for ℓ = 15 μm down to 3 μm the agreement is excellent but again there are significant differences for the shorter sampling interval of ℓ = 1 μm. It will be observed from the magnitudes of the curvatures shown in figures 4-13(a), (b), (c), (d), (e), that as the sampling interval is decreased one is concerned with asperities of smaller and smaller radius. This is made quite clear in figure 4-14 which compares theoretical values of the mean curvature of the peaks with the values found from the profiles using differing sampling intervals. Once more the only significant divergence between theory and experiment occurs at the shortest sampling interval (ℓ = 1 μm).

The results obtained at the shorter sampling intervals suggest

Figure 14-13(a). Peak curvature distribution correlation 0.10.

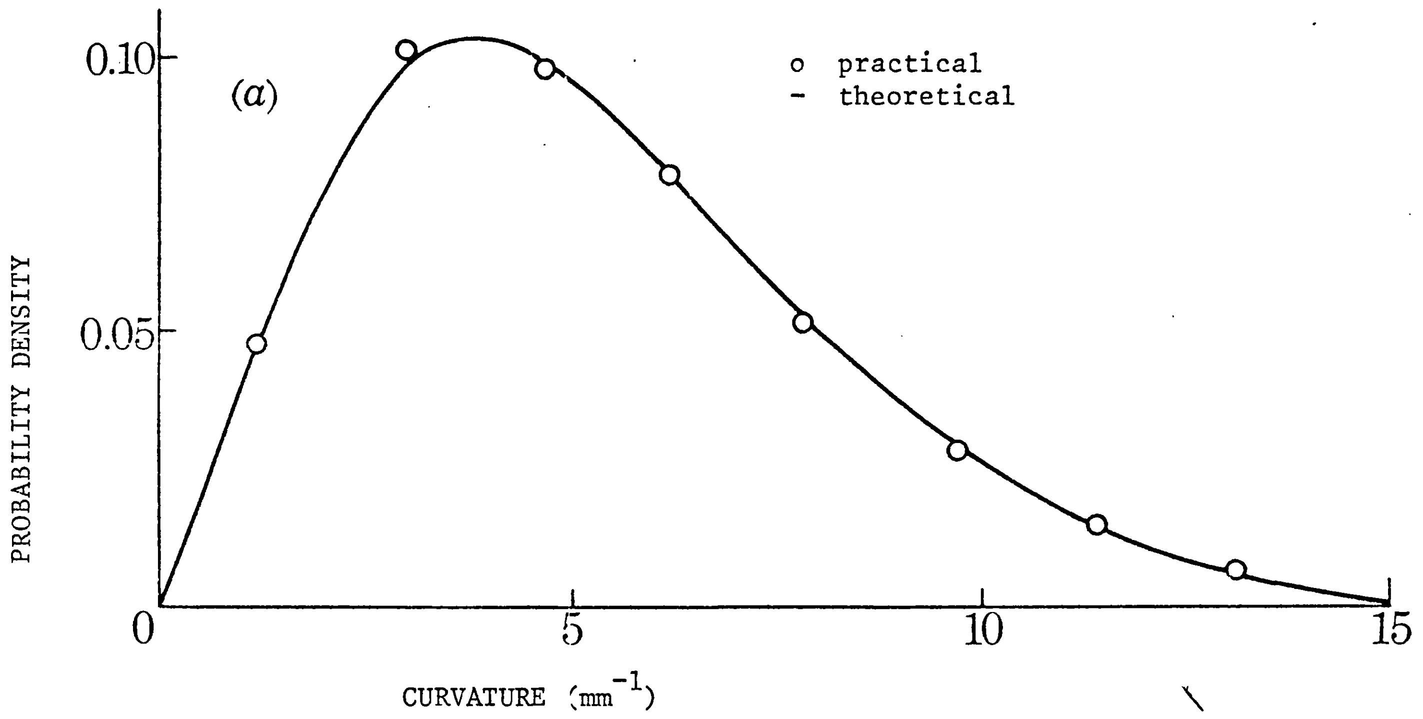


FIG 4-13b

PEAK CURVATURE DISTRIBUTION CORRELATION 0·40

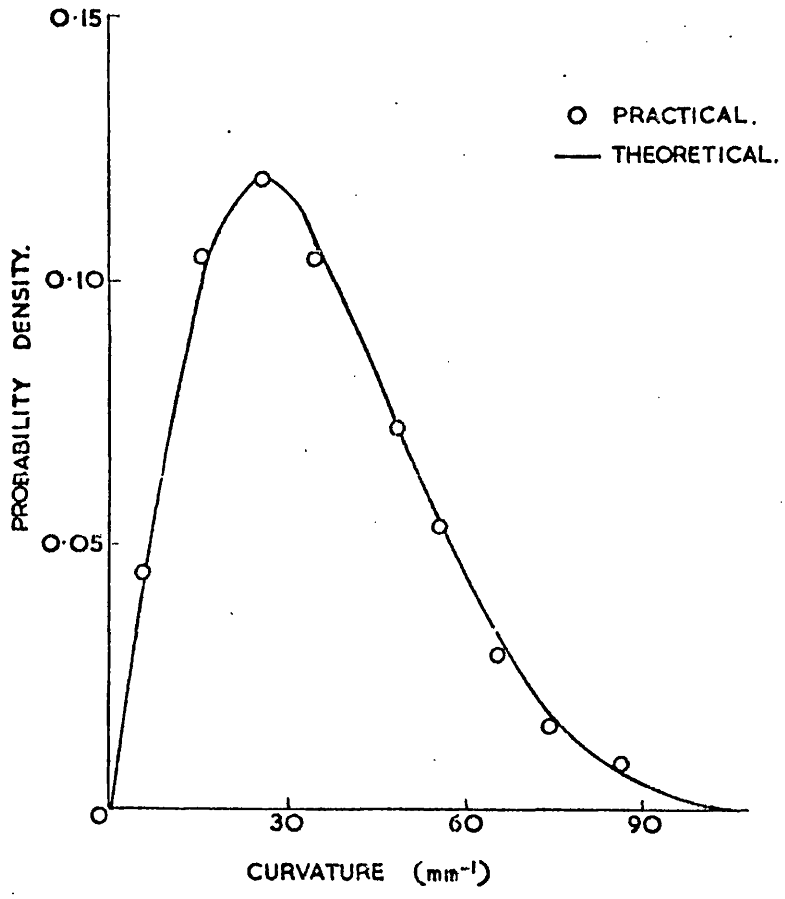


Figure 4-13(c). Peak curvature distribution correlation 0.63

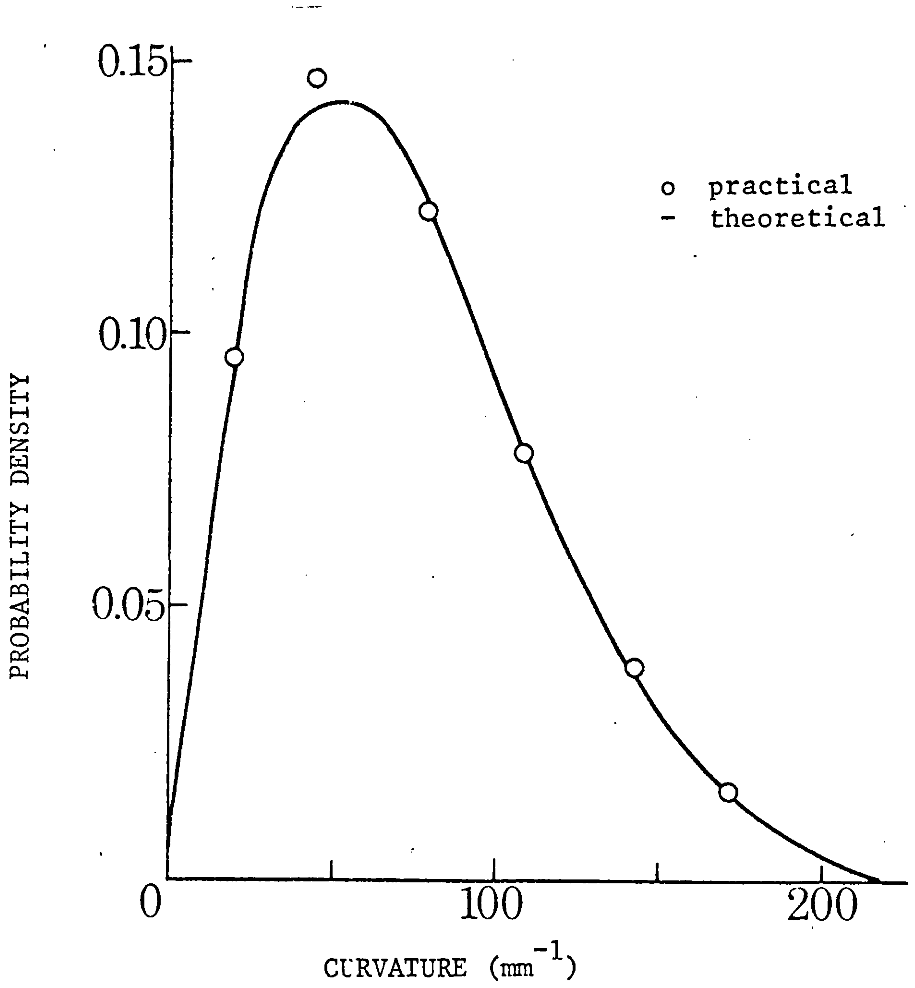


FIG. 4-13d

PEAK CURVATURE DISTRIBUTION CORRELATION 0·74.

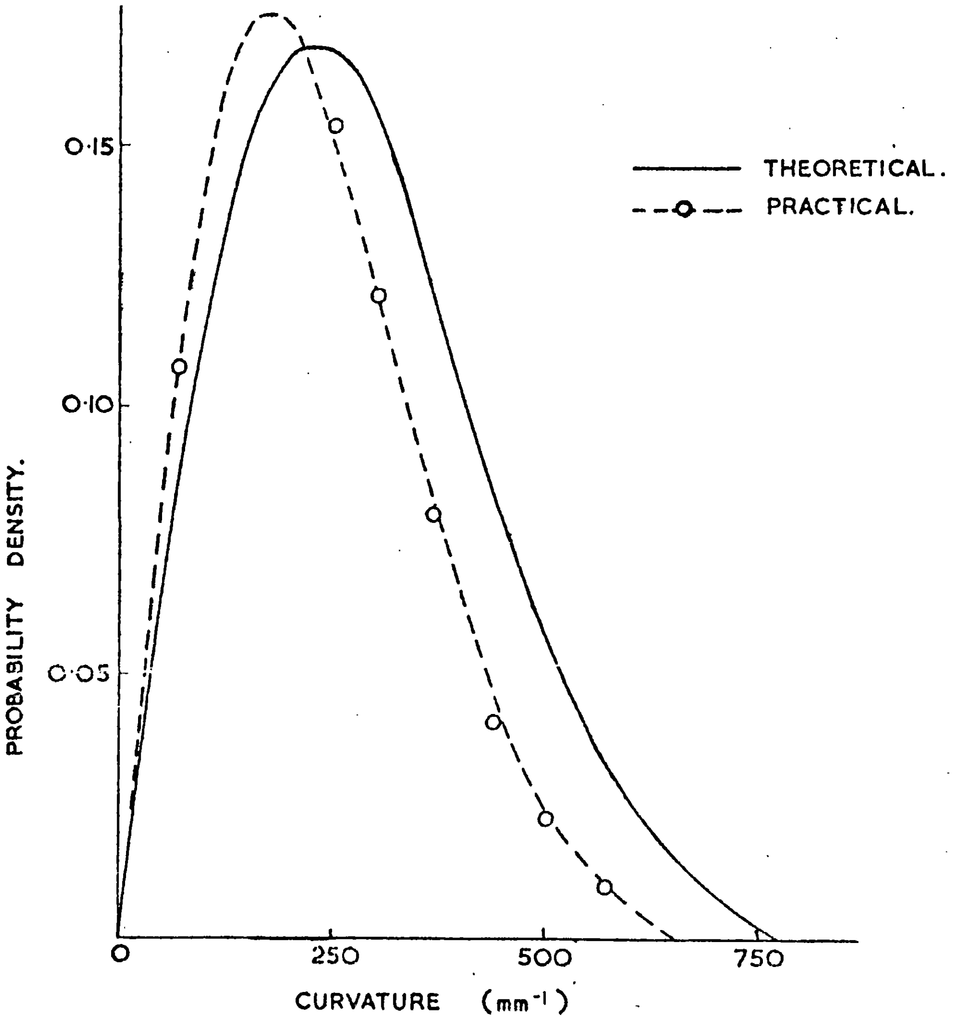


Figure 4-13(e) Peak curvature distribution correlation 0.86

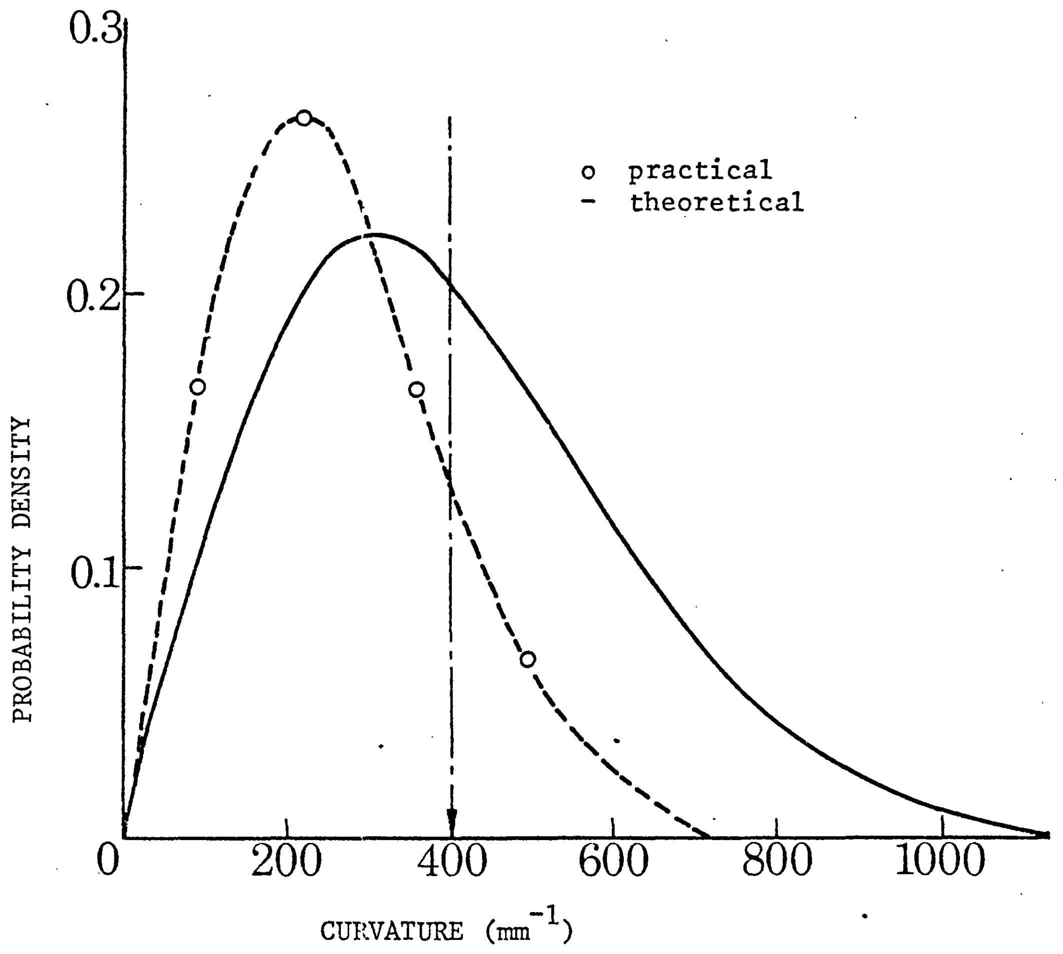
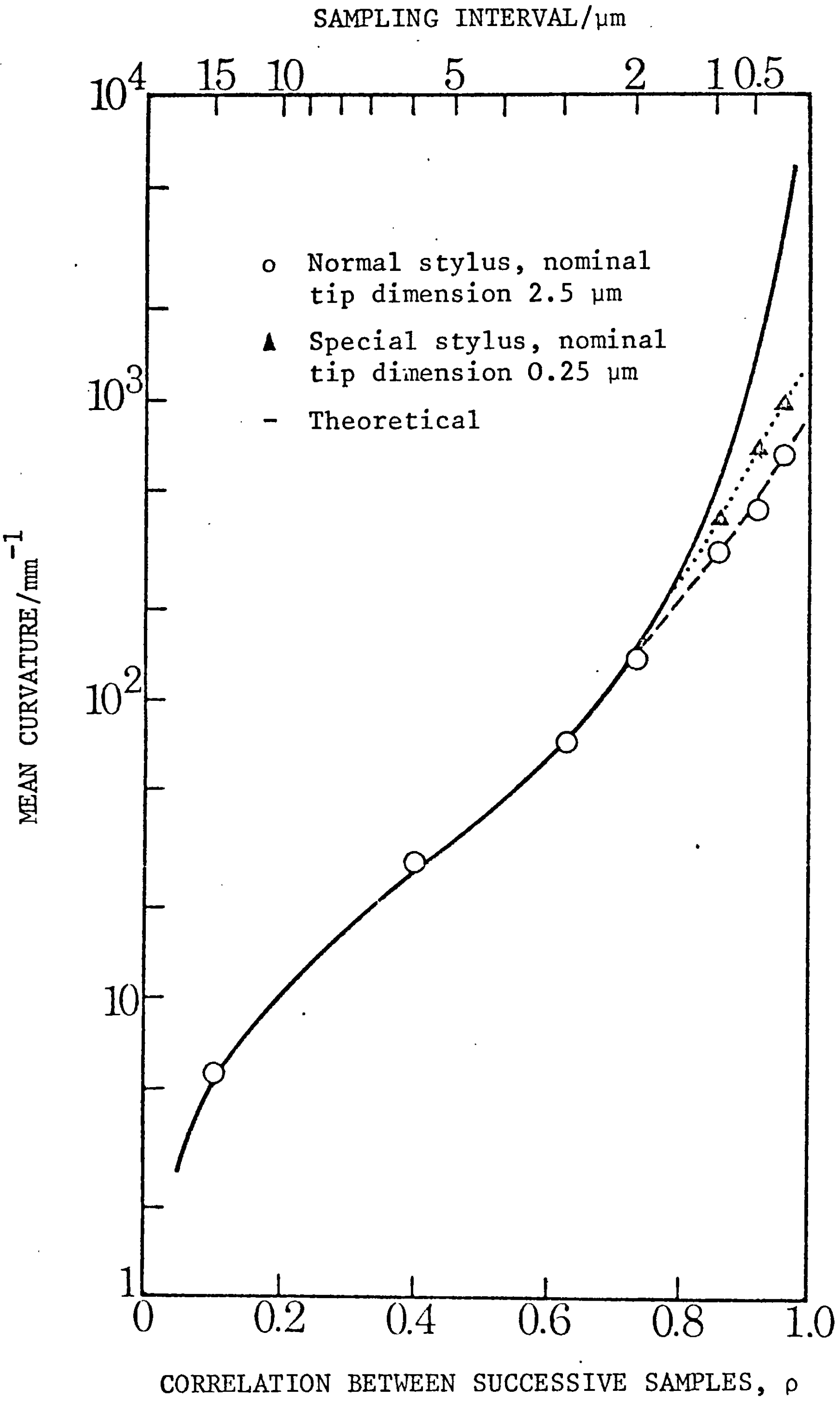


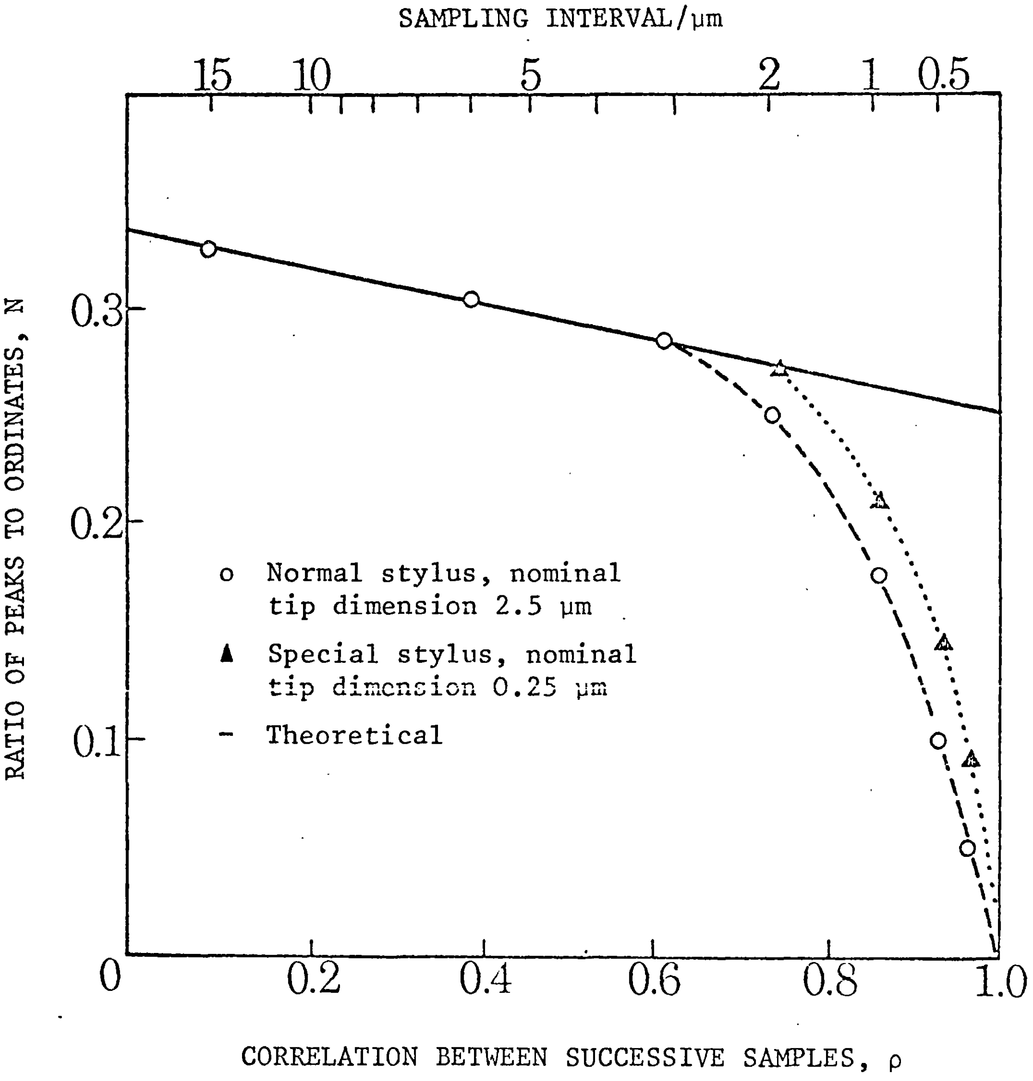
Figure 4-14. Mean curvature of the peaks as a function of the correlation (ρ) between successive samples. The full line gives the theory. The experimental points are derived from digital analysis of profiles from Aachen 64-13 (σ = 0.5 μm, β* = 6.5 μm).



that the measurement of the surface profiles are affected by the finite size of the stylus. In figure 4-13(e) a value of the nominal stylus curvature has been indicated, this is taken as the reciprocal of the nominal tip dimension of the stylus. The character of the divergence between theory and experiment shown in figure 4-13(c) is certainly consistent with the assumption that it arises from the finite size of the stylus. The total number of peaks detected is less than that forecast by the theory and the distribution has apparently been distorted towards smaller values of the curvature.

It is, of course, equally possible that the surface used in this work does not conform at these shorter wavelengths to the model assumed here. Figures 4-11(e) and 13(e) would then imply that the structure of shorter wavelengths, although present in the model, does not exist upon the surface. In an attempt to resolve this question experiments were performed with a stylus having a smaller tip dimension. A discussion of the experimental procedure and the special problems involved has been given in Chapter 3. The results are shown in figure 4-15 where the ratio (N) of peaks to ordinates is plotted against the correlation (ρ) between successive samples. It will be recalled that the theory (equation (4.32)) forecasts that this ratio varies between 0.33 ($\rho = 0$) and 0.25 ($\rho = 1$). Figure 4-15 shows once more a divergence between theory and experiment for sampling intervals of less than 2 μm; in this region the number of peaks detected falls well below the theoretical values. Similar plots showing a decline have been presented by Sharman (1967), but no explanation of the cause was

Figure 4-15. Ratio (N) of peaks to ordinates as a function of the correlation (ρ) between successive samples. The full line gives the theory. The experimental points are derived from digital analysis of profiles from Aachen 64-13.



advanced. Figure 4-15 also shows that when using a stylus with a smaller tip dimension the decline is delayed to smaller values of the sampling intervals. Clearly, therefore, stylus resolution is a significant factor in the behaviour in this region. This is clearly brought out by looking not at just the distribution of the peaks but also at the distribution of the valleys.

That the stylus resolution is important in the valley distribution as well as the peak distribution can be seen in figure 4-16. This shows a part of a graph taken on Aachen 64-13 with no magnification difference between the horizontal scale and the vertical scale. The magnification of 2000 that was chosen is necessary for two reasons; (a) the recorder would not see the possible sharp changes in the slope at the bottom of the valleys if the magnification was lower in the horizontal direction, and (b) the surface is fine enough in roughness to require a reasonable magnification vertically. It can be seen that the valleys appear to be sharper than the peaks. This is due to the stylus tip and is not simply a characteristic of the surface.

It is a straightforward exercise to check that this is so. This is done by making an epoxy replica of the surface and retracking in the same way; using a replica turns the peaks into valleys and vice-versa. The same effect was noticed - the apparent valleys were still sharper.

The reason for this apparent curvature discrimination of the stylus against peaks rather than valleys is because the stylus is in

FIG. 4-16

AACHEN 64-13 TAKEN WITH 2·5 μm STYLUS

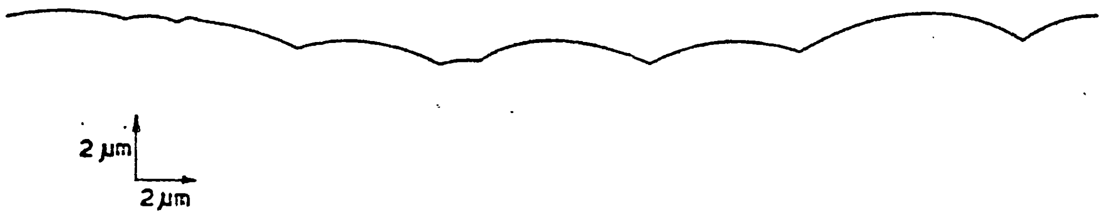


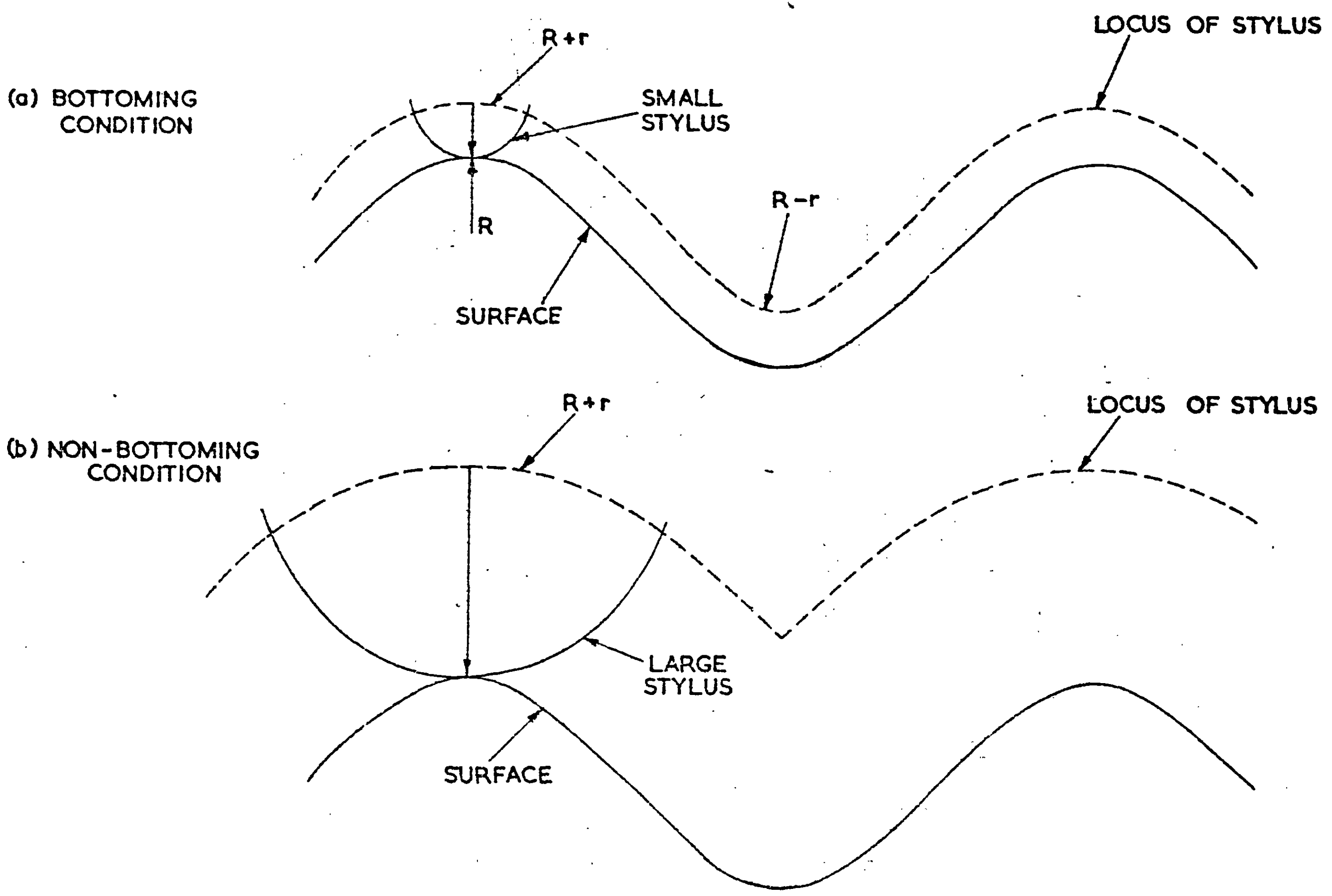
TABLE 4.3

Comparison of some peak and valley results for 0.25 μm sampling tip of stylus (μm).

Tip of stylus (μm)	Average curvature of peak (mm^{-1})	Average curvature cf valleys (mm-1)	Ratio peaks/ ordinates	Ratio valleys/ ordinates
2.5 x 2.5	400	900	0.04	0.06
0.25 x 0.25	700	1000	0.082	0.085

FIG. 4 –17

EFFECT OF STYLUS RESOLUTION



effect a non-linear filter. This is brought about because when a stylus runs over a peak the effective radius of curvature of the movement of the stylus is equal to the radius of the peak, plus the radius or equivalent radius of the stylus. When the stylus runs into a valley one of two things happens,(a) the stylus runs in and out of the bottom of the valley, or (b) the stylus does not bottom. In case (a) the radius of curvature of the locus of the stylus is that of the valley bottom minus the radius of the stylus. In the figure a rounded stylus has been shown for clarity. In case (b) where the stylus does not bottom there is a discontinuity in the slope of the locus because the stylus has effectively jammed in between the valley sides, figure 4-17(b). From both of these cases it is clear that the valleys are sharper than the peaks - as revealed by a stylus instrument. This fact has been noticed in the digital analysis of the information. Some results are shown in Table 4-3. It shows the difference up clearly.

Another consequence of the difference in sharpness of the peaks and valleys can be seen in the ratio of the number of peaks and valleys to the number of ordinates shown in Table 4-3. There appear especially for the big stylus to be more valleys than peaks. This is a result of four things: the big stylus, the quantisation interval, the sampling interval and the three-point definition of a peak or valley. Because the stylus is large the peak appears blunt, this in turn means that for the short sampling interval there is little height difference between successive ordinates; this difference in some instances will be smaller than the quantisation interval for a few ordinates near to a peak where the changes in level are small anyway, and because of the three-point definition no peak is registered. The effect will tend to be the opposite for valleys.

From the results obtained in this section it is clear that the stylus resolution does have a significant effect on the results of a digital analysis of a surface profile especially for the case of a small sampling interval.

4.4 Some applications of the model

4.4.1 Extension to three dimensions - definition of a summit

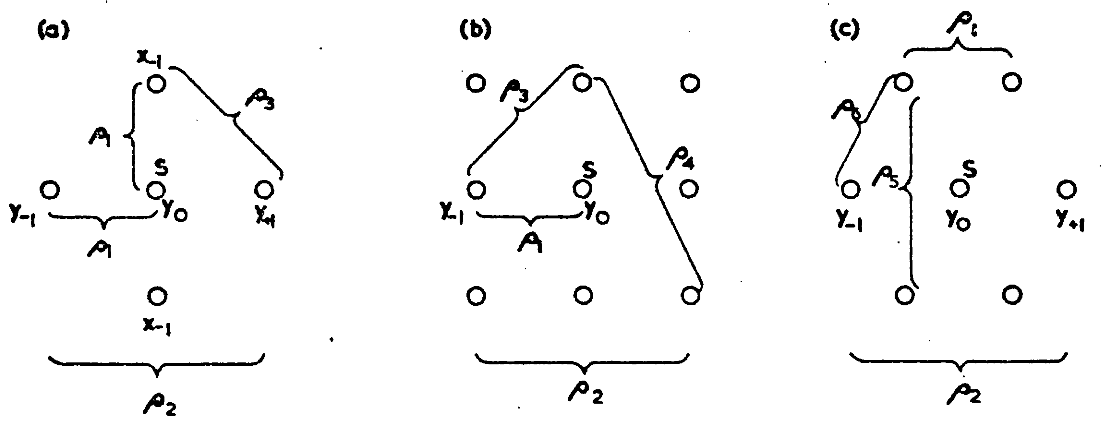
In Section 4.3.2 a peak was defined by reference to selected points on a profile. Thus a peak was defined as a high point on a two-dimensional representation of the surface. Similarly we shall now consider a summit as a high point on a three-dimensional representation of the surface. A peak was simply represented by three-point analysis. We now consider the corresponding representation of a summit.

Obviously in engineering terms a summit or strictly three summits would automatically be defined if a plate were to be rested on a surface. If the plate is made larger, to cover more of the surface then the same three summits would not likely be those that contacted in the first case. The definition considered here is that for the first contact. As before the surface will be

FIG. 4—13

DEFINITION OF SUMMIT – ORDINATE CONFIGURATIONS

(a) 5 ORDINATE MODEL
(b) 9 ORDINATE MODEL
(c) 7 ORDINATE MODEL



considered to be made up of individual elements which may be the sampled digital values. Figure 4-18(a) shows that the minimum convenient number of ordinates required for a definition is five; one in the centre being higher than four, each of equal weight, on the four sides. In this picture the surface is assumed to be isotropic so that these four ordinates are all the same distance from the central one. This definition corresponds to a summit which is defined when a circular plate of radius equal to unit ordinate spacing is laid on the surface. If the radius taken is $\sqrt{2}$ times the ordinate spacing the number of ordinates now required is 9. However, in this case, the ordinates in the definition are not equal weight relative to the central ordinate (figure 4-18(b)).

Another simple justification for the five ordinate model is that the equation of a summit of height z_0 having an apex at $x = 0$, $y = 0$ is $z = z_0 - a_1x^2 - a_2y^2 - 2a_3xy$ which has four unknowns, consequently at least four points need to be available in order to define it. For this purpose only the five point model needs to be used.

To show how difficult it is to define a summit even in these simple ways consider figure 4-18(c). Just by changing the ordinate configuration the minimum number of ordinates in the definition changes from 5 to 7; the six ordinates surrounding the central one S being of very nearly equal weight. This suggests that the number of summits likely to be found on a surface by any digital means is likely to be variable depending on the method of definition.

Take the simplest case, assuming isotropic geometry and denoting the ordinates orthogonal to the y values as x values, the configuration defining a peak is shown in figure 4-18(a).

The co-variance matrix M in the case of a general correlation function is given in the following order y_{+1}, x_{+1}, y_0, x_{-1}, y_{-1}

$$M = \begin{vmatrix} E(y_{+1}\cdot y_{+1}), & E(y_{+1}\cdot x_{+1}), & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & E(y_{-1}\cdot y_{-1}) \end{vmatrix} \qquad (4\text{-}62)$$

$$M = \rho_1 \begin{vmatrix} 1 & \rho_3 & \rho_1 & \rho_3 & \rho_2 \\ \rho_3 & 1 & \rho_1 & \rho_2 & \rho_3 \\ 1 & 1 & 1 & 1 & 1 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & \rho_1 & \rho_3 & 1 \end{vmatrix} \qquad (4\text{-}63)$$

When the autocorrelation function is exponential

$$\rho_2 = \rho_1^2 \text{ and } \rho_3 = \rho_1^{\sqrt{2}} \qquad (4\text{-}64)$$

For the case when the ordinates can be considered independent $\rho_1 \sim 0$ and the joint probability density $f(y_{+1}, x_{+1}, y_0, x_{-1}, y_{-1})$

becomes

$$f(y_{+1},x_{+1},y_0,x_{-1},y_{-1}) = \frac{1}{(2\pi)^{5/2}}$$

$$\exp\left(-\frac{1}{2}(y_{+1}^2+x_{+1}^2 +y_0^2+x_{-1}^2+y_{-1}^2)\right)$$

(4-65)

which gives the probability of a summit at y, $f^o(y)$ to be given by

$$f^o(y) = \frac{1}{16\sqrt{2\pi}} \left[1 + \text{erf }(y/\sqrt{2})\right]^4$$

$$\exp\ (-\ y^2/2 = \frac{1}{\sqrt{2\pi}}\ \Phi^4(y)\ \exp\ (-\ y^2/2)$$

(4-66)

where the circle over the letter f indicates a summit property as distinct from an asterisk for a peak property.

The case where the ordinates are correlated is clearly much more complicated. We shall, therefore, make a number of gross simplifications to obtain some feel of the probable solution.

For the case when ρ_1 is large compared with ρ_3, then the equation (4-66) becomes

$$f^o\ (y,\rho) = \frac{1}{16\sqrt{2\pi}} \left[1 + \text{erf }(y/\sqrt{2}\ \sqrt{\frac{1-\rho}{1+\rho}}\ )\right]^4$$

$$\exp\ (-y^2/2)\ f^o\ (y,\rho) = \frac{1}{\sqrt{2\pi}}\ \Phi^4\ \left(y\ \sqrt{\frac{1-p}{1+\rho}}\ \right)$$

(4-67)

To convert equations (4-66) and (4-67) to true probability densities require that they are divided by the ratio of summits to ordinates in both cases. In the case of equation (4-66) this ratio is one fifth and for equation (4-67) it is one sixteenth.

If nine ordinates had been used for the definition then for the case of independent events

$$f^{o}\ (y) = \frac{1}{256\sqrt{2\pi}} \left[1 + \text{erf}\ (y/\sqrt{2})\right]^{8}$$

$$\exp\ (-y^{2}/2)\ f^{o}\ (y) = \frac{1}{\sqrt{2\pi}}\ \Phi^{8}\ (y)$$

(4-68)

The ratio of summits to ordinates is one ninth in this case. Making again the not inconsiderable assumption that for the correlated case $\rho_1 >> \rho_3$ is large compared with the other correlations between all ordinates other than the central one, a situation likely to be most true for values of correlation remote from ~~between~~ zero and high correlation, then

$$f^{o}\ (y,\rho) = \frac{1}{256\sqrt{2\pi}} \left[1 + \text{erf}\ (y/\sqrt{2}\sqrt{\frac{1-\rho}{1+\rho}}\right]^{8}$$

$$\exp\ (-y^{2}/2) = \frac{1}{\sqrt{2\pi}}\ \Phi^{8}\ (y\sqrt{\frac{1-\rho}{1+\rho}}\ )$$

(4-69)

from which it is clear that under these circumstances and with these assumptions the ratio of summits to ordinates would be $\frac{1}{256}$. Consequently assuming that equations (4-66) to (4-69) represent the sort of spread of sampling interval and definition that can be expected then the variation in the number of summits found over a given area could be very high; the actual value depending on the sampling interval and the definition. Just how valid the assumptions are that were made in the high correlation case will be discussed in Section 4.5.

Another point concerning the assessment of the number of summits as a fraction of the number of ordinates occurs when trying to predict the number of summits from the number of peaks in the two dimensional case. If the ordinates are independent of each other and one is considering three-point analysis, the probability of one ordinate being a peak is 1/3. Simply squaring this to give the probability that this ordinate is a summit implies that $f^* (y_{+1}, y_0, y_{-1})$ is independent of $f^* (x_{+1}, y_0, x_{-1})$ which it obviously is not. The reason for this is that if y_0 is a peak in one direction, this fact means that, on average, y_0 lies above the centre line; it therefore has a greater probability than normal of also being a peak in the other direction. One could carry out an experiment in which an area of surface is completely covered with a grid of data logged traces. Scanning the surface in the y direction, the number of peaks in every trace could be counted and the total number of peaks in the area found. A similar procedure could be followed by scanning in the x direction with a similar result if the surface is isotropic. But the ratio of summits to ordinates is not simply the square of the ratio of peaks to ordinates found in one scan. In our second scan we wish to find the number of summits; that is the number of points, previously defined as peaks, which are peaks a second time. For the independence case discussed above, this ratio would be 3/5. Hence the number of summits in an area cannot be determined directly from one digital trace, the only really valid way to deal with summits is by using the multi-normal distribution in the three-dimensional case.

To summarise. The correct procedure for estimating digitally the ratio of summits to ordinates is to work from the three-dimensional definition and this should preferably be the five-point definition because it is the logical extension of the three-point definition of a peak; like that definition it uses the minimum amount of data. Moreover all of the outer points have equal weight in the definition.

4.4.2 Variability and confidence limits

The importance of β^* in the specification of surface finish has been stressed but it has additional significance in its own right. Consider the measurement of the R_a or RMS roughness value of a random type of manufactured surface; it is often desired to know the confidence limits of such a measurement and this is easily expressed if one knows the standard deviation of a large number of such measurements made upon the same surface. Alternatively this can be estimated if β^* is known. It can be shown that the standard deviation of such a measurement of the R_a roughness of a random

FIG. 4 – 19

EFFECT OF STYLUS RESOLUTION.

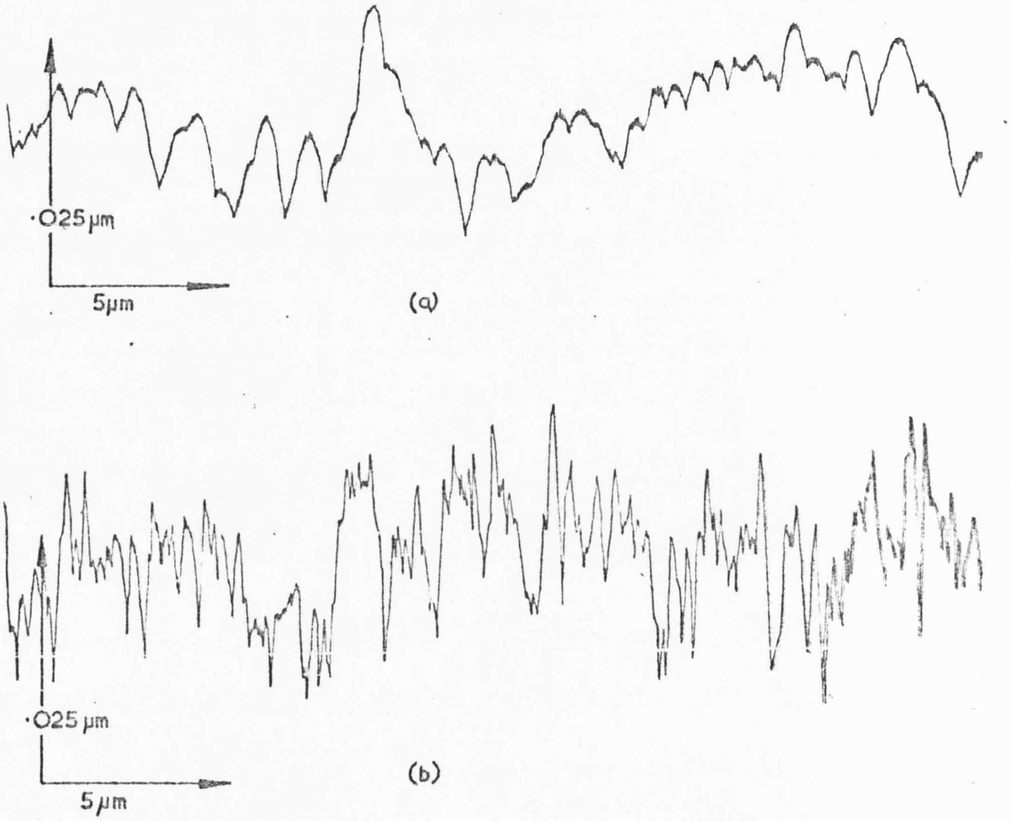


CHART (a) IS A PROFILE OF A SMALL SECTION ON A FINELY LAPPED GAUGE BLOCK

CHART (b) SHOWS A SECTION OF THIS SAME SURFACE AS MEASURED BY THE SHARP STYLUS surface, when normalised as a ratio of the mean value is approximately

$$\approx \quad 1/\sqrt{2\underline{M}} \qquad \text{(4-70)}$$

In this equation $\underline{M}$ is the ratio of the assessment length used in the measurement of R_a to the distance between points on the surface (2.3β*) which just provides effectively independent events. (This ratio $\underline{M}$ is analogous to the bandwidth-duration product used in communications theory to estimate the reliability of data, Bendat and Piersol (1966)).

For example on the 8 mm cut-off range, the Talysurf 4 instrument has an assessment length of 3.8 mm. The value of 2.3β* for Aachen 64-13 is 15 µm. Thus the normalised value of the standard deviation of R_a readings on this surface should be 4.5%. A measured value of the standard deviation for Aachen 64-13 based on a large number of readings was 4.3%.

Knowledge of this, the effective number of degrees of freedom of the waveform, enables the variations in other parameters to be estimated from a limited length of profile, for instance, the RMS value, the variation in which again has a formula similar to (4-70). This is derived from a knowledge of the variance in the mean square estimation of the signal.

$$\text{Var.(Mean Square)} = \frac{2}{L} \int_{-L}^{L} (1 - \frac{|\beta|}{L}) \left| (C(\beta))^2 + 2\underline{\mu}^2 \, C(\beta) \right| d\beta$$

$$\text{(4-71)}$$

FIG. 4 –20

RELATIONSHIP BETWEEN AREA OF CONTACT AND LOAD

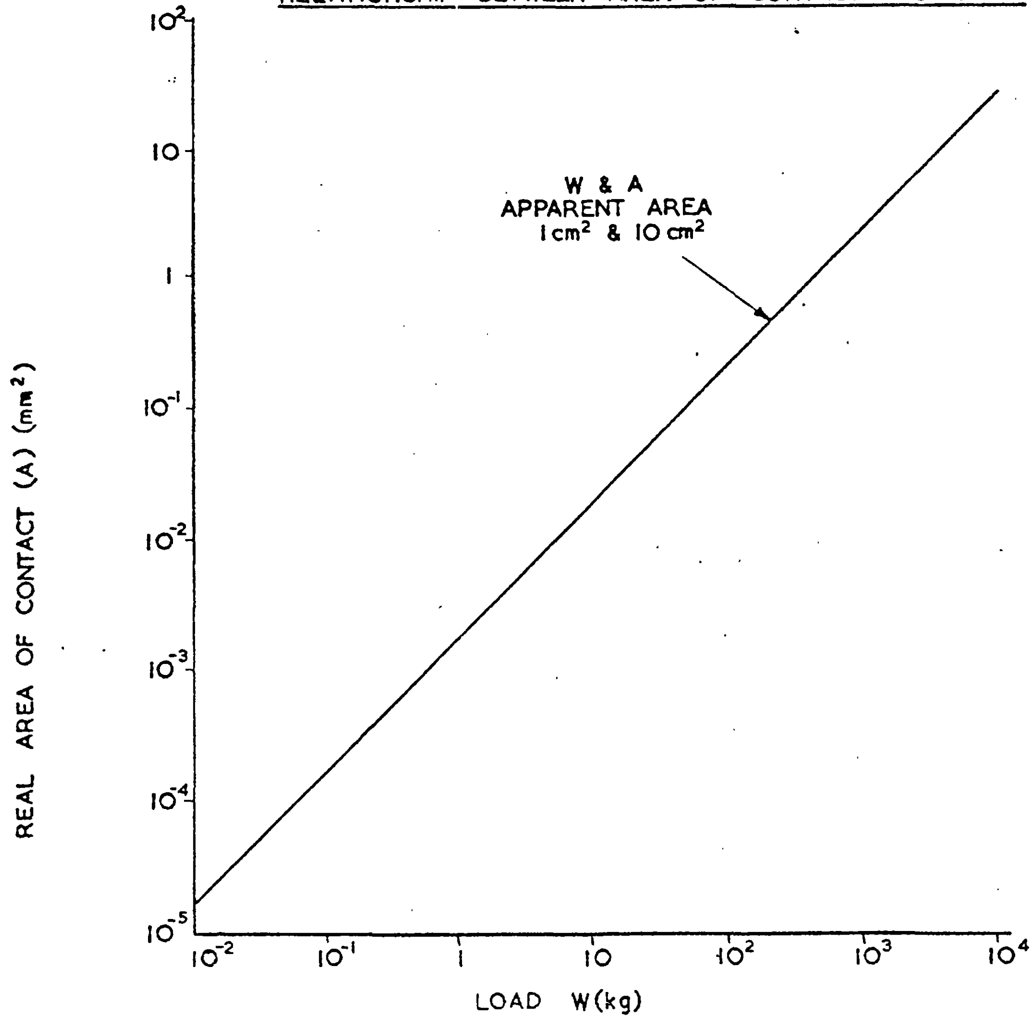
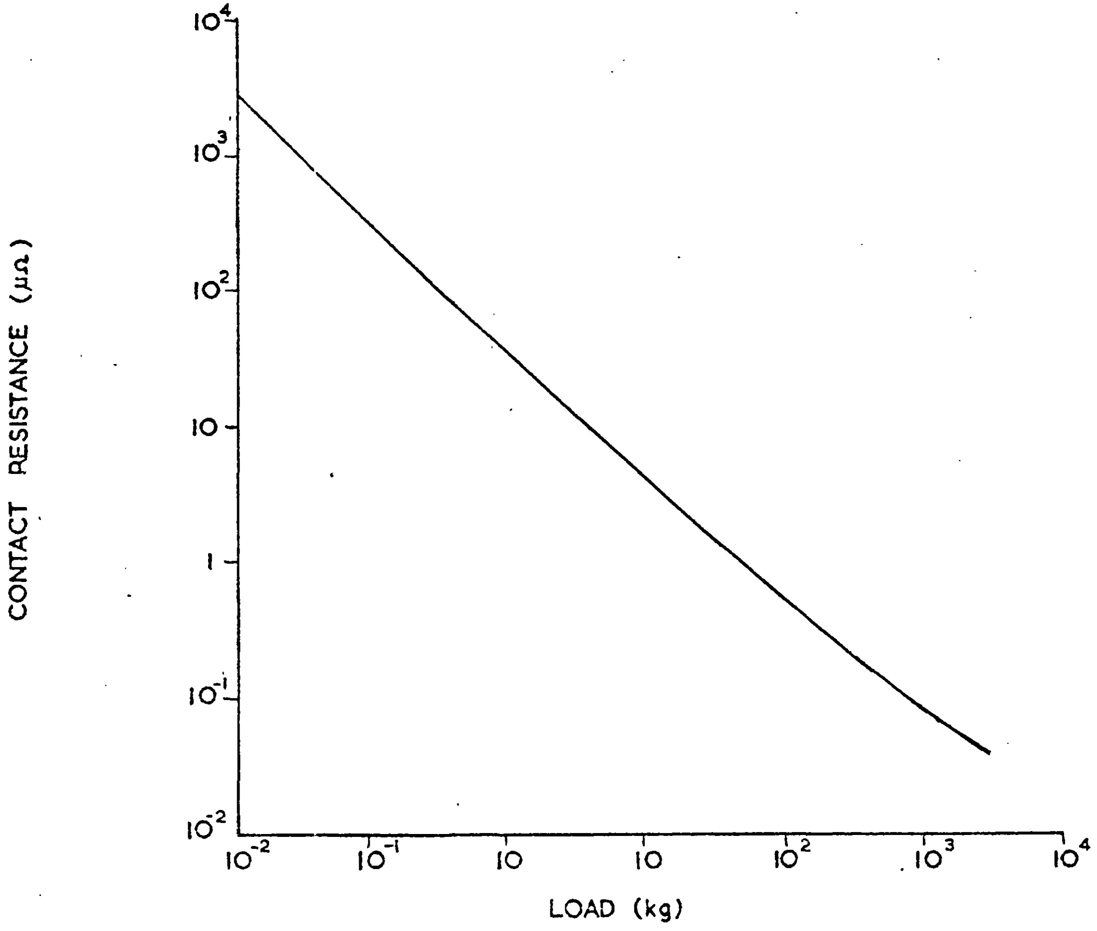


FIG. 4–21

RELATIONSHIP BETWEEN CONTACT RESISTANCE AND LOAD



where C (β) is the autocorrelation and μ is the mean value of the waveform (Bendat and Piersol 1966).

4.4.3 Application of the model to theories of surface contact

In this section the relationship between separation of bodies, load, area of contact and conductance under elastic deformation will be discussed to provide a physical background to the application of models of surface asperities.

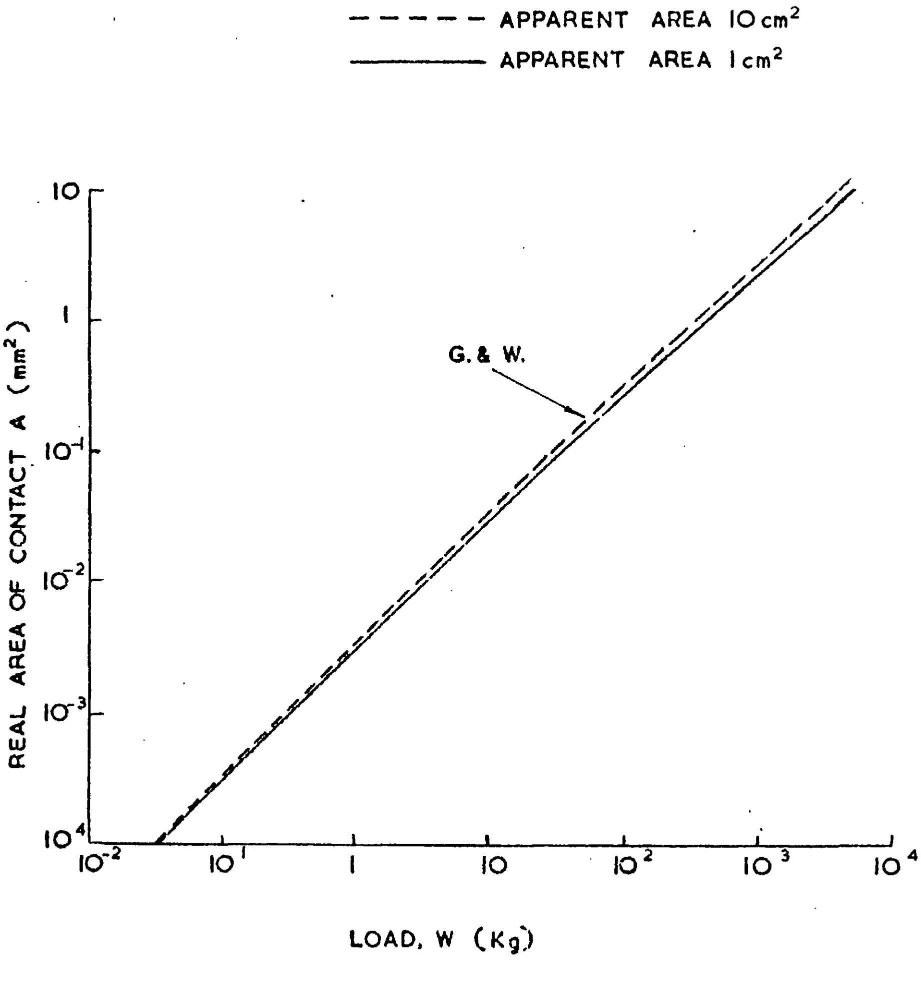
In Chapter 5 the contact of bodies will be considered taking the size and shape of the upper specimen into account, however, a similar approach to that of Greenwood and Williamson will be taken in the present section. This latter part has been carried out by Mr. R. A. Onions and will form part of his Ph.D., thesis, the following is a summary of his results.

As in the Greenwood and Williamson treatment it will be assumed that the peaks are independent and do not interact upon deformation. The contact between a non-deformable flat surface and a deformable surface, having the characteristics of the model of this chapter, will be assumed. As in most models of surface contact the dependent variables (Area, Load, Contact resistance) are calculated as functions of one independent variable (the separation).

In this approach the surface is considered to be made up of asperities made up of spherical caps of radius R.

FIG. 4-22

RELATIONSHIP BETWEEN REAL AREA AND LOAD.



Consider the contact of one of those asperities, the area of contact δA is given by $\delta A = \pi R\omega$ using Hertz's theory where $\underline{\omega}$ is the compliance.

For two surfaces whose mean levels are separated by k, at any level y, only a ~~proportion~~ of peaks $f^*(y,C)$ equation (4-47) will have a curvature C. Hence the proportion of real area of contact due to peaks at this height having this curvature $(\delta A)_{y,C}$ is given by

$$(\delta A)_{y,C} = \frac{\pi}{C} \, \omega \, f^*(y,C) \qquad (4\text{-}72)$$

where in this formula it is assumed that only the independent peaks are to be considered, hence the use of $f^*(y,C)$ rather than $f^*(y,\rho,C)$. For all peaks at this height having any curvature equation (4-72) has to be integrated with respect to C

$$\text{mean } (\delta A)_y = \pi(2.3\beta^*)^2 \, (y-k) \int_0^\infty \frac{f^*(y,C)}{C} \, dC \qquad (4\text{-}73)$$

since $R = \dfrac{(2.3\ \beta^*)^2}{\sigma C}$ for independence.

For all peaks contributing to the area A lying between k and ∞

$$\text{mean } (\delta A) = \pi(2.3\beta^*)^2 \int_k^\infty (y-k) \int_0^\infty \frac{f^*(y,C)}{C} \, dC \, dy \qquad (4\text{-}74)$$

FIG. 4-23

RELATIONSHIP BETWEEN PRESSURE AND SEPARATION.

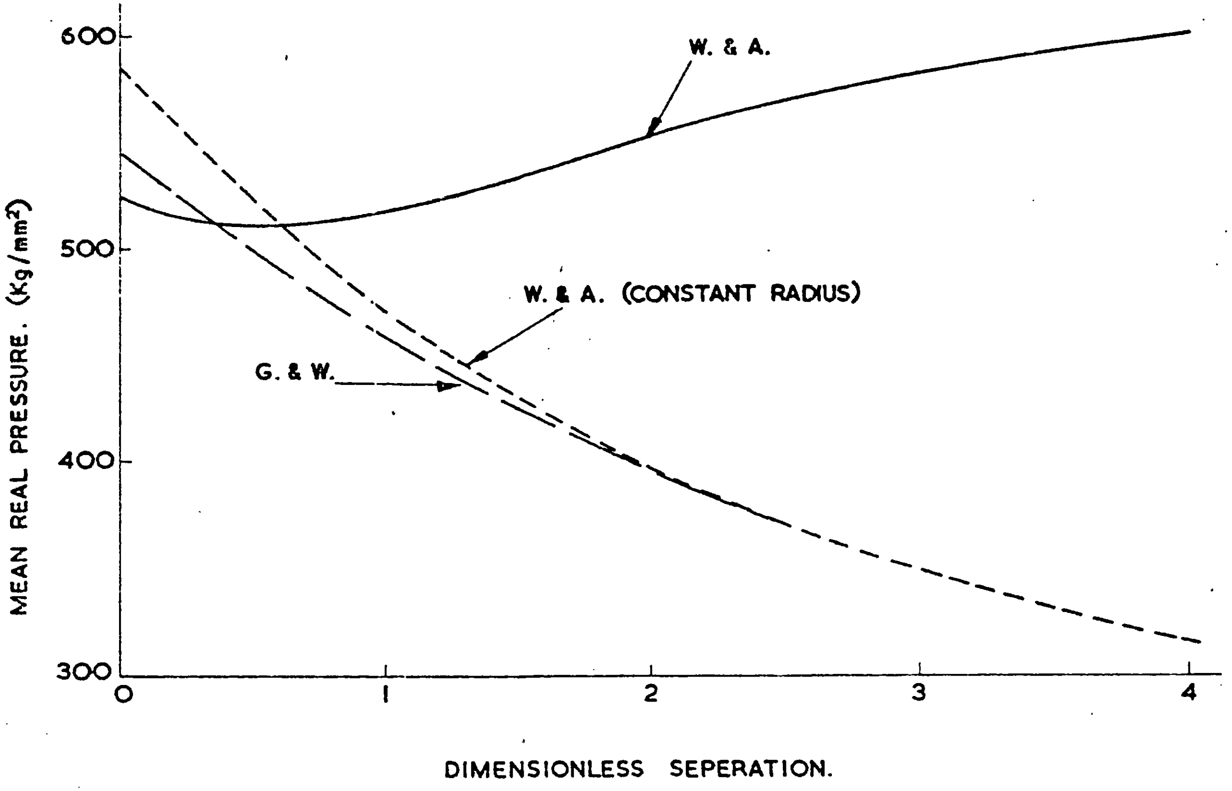
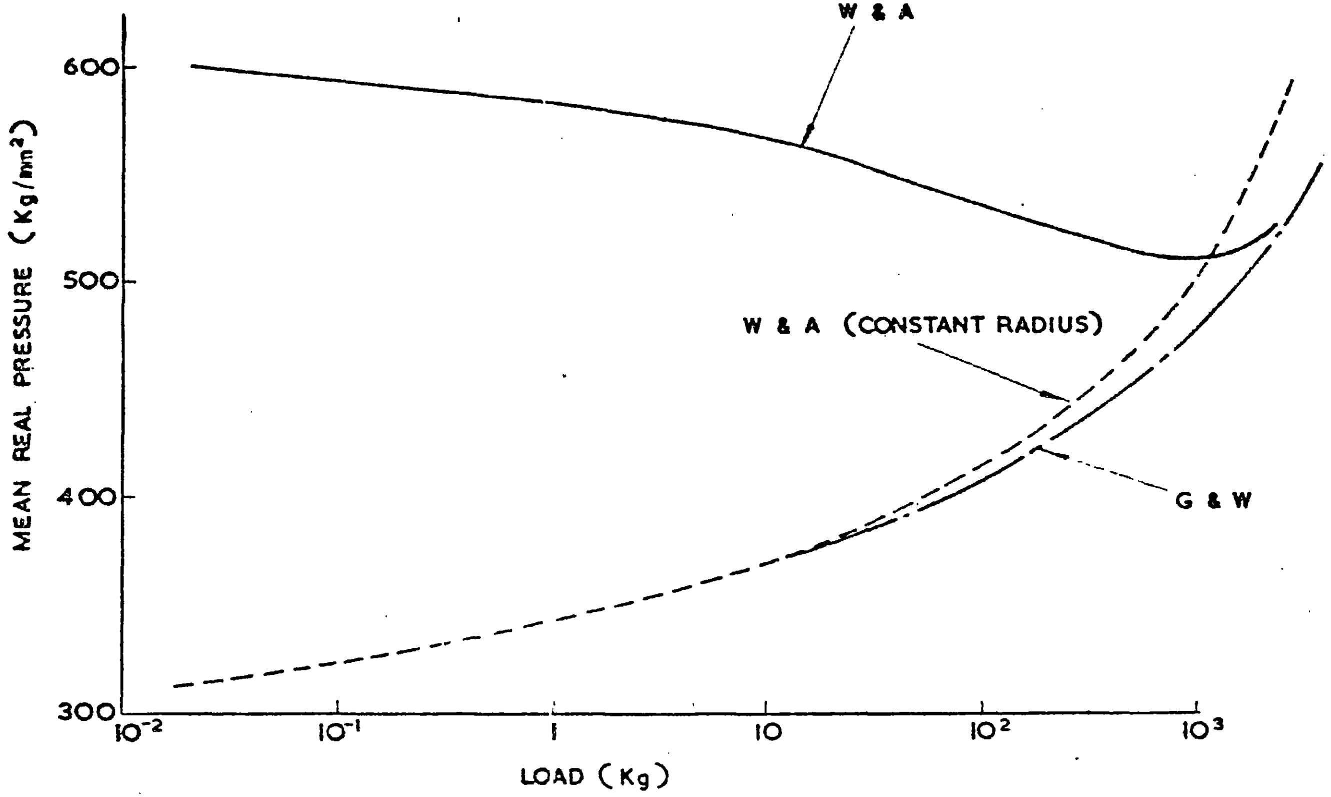


FIG. 4-24

RELATIONSHIP BETWEEN PRESSURE AND LOAD.



hence the total area of contact A is given by

$$A = \pi(2.3\beta^*)^2 \; N \int_k^\infty (y-k) \int_0^\infty \frac{f^*(y,c)}{C} \, dC \, dy \qquad (4\text{-}75)$$

where N is the number of asperities per unit area estimated at the mean line from β^{*2}.

Similarly expressions for the total load W and the conductance (Holm 1958) may be obtained.

$$W = \frac{N\sqrt{2}}{3\pi} \, E' \, (2.3\beta^*)\sigma \int_k^\infty (y-k)^{3/2} \int_0^\infty \frac{f^*(y,C)}{\sqrt{C}} \, dC \; dy$$

$$(4\text{-}76)$$

and $$G = \frac{N(2.3\beta^*)}{\underline{r} \; \pi \; \sqrt{2}} \int_k^\infty (y-k)^{1/2} \int_0^\infty \frac{f^*(y,C)}{\sqrt{C}} \, dC \, dy \qquad (4\text{-}77)$$

where $\underline{r}$ is the resistivity.

Using these expressions, graphs can be plotted showing the variation of real area of contact with load (figure 4-20), contact resistance with load (figure 4-21), variation of mean real area with dimensionless separation (figure 4-22) and variation of mean pressure with dimensionless separation and load (figures 4-23 and 4-24). These graphs will be discussed in Section 4.5.3 together with the implications of the differences when compared to those obtained by Greenwood and Williamson.

4.5 Discussion

4.5.1 Digital analysis

The points associated with three-point analysis and the loss of information inherent in its use have been mentioned, together with the other difficulties involved in its implicit assumptions. The main point at issue in this sort of approach is to decide the range of sampling intervals from which it is acceptable to use information in devising theories for surface contact.

The significance of the shorter wavelength structures (revealed by the use of the shorter sampling intervals) in the contact and rubbing of surfaces may be questioned; in any event, a lower limit to the acceptable range is set by stylus resolution. At the other extreme it is clear that the use of very long sampling intervals will give results which have little physical significance. A more relevant spacing of ordinates is that which will define the dominant or main structure of the profile. This spacing is that which just makes successive ordinates independent of each other; this corresponds to the zero order Markov sequence equation (4-20). How these ordinates have to be spaced for independence depends on the type of autocorrelation function, see Section 4.2.1. Dependent on the arguments then put forward as a general guide to the main structure of relevance in contact problems a sampling interval of $2.3\beta^*$ has been used.

There are a number of ways in which to justify the spacing of events at 2.3β*, none of them precise. However, estimates can be made by considering different ways of defining the main wavelengths. For independent events the main structure of the waveform has an equivalent wavelength of about 10β*. This can be seen by using a simple Bernouilli argument. Let the profile have two states; the one corresponding to when the profile is above the mean line (state 0) and the other (state 1) when the profile is below the mean line. For there to be a crossover and back, corresponding to the equivalent of about half a wavelength, then there must be a change of state - equivalent to a transition of say p_{01} followed some time later by p_{10}. During this time only p_{11} transitions would be allowed.

Hence the probability of half a wavelength happening in two adjacent intervals is $p_{01} \cdot p_{10}$; for three intervals it is $p_{01} \cdot p_{11} \cdot p_{10}$ and for m intervals it is $p_{01}\, p_{11}^{m-2} \cdot p_{10}$. Hence the average number of intervals $\bar{n}$ over which there has been a transition across the mean line and back is given by the mean value of the distribution. Thus

$$\bar{n} = \frac{\sum_{i=2}^{\infty} i\, p_{01} \cdot p_{11}^{i-2} \cdot p_{10}}{\sum_{i=2}^{\infty} p_{01} \cdot p_{11}^{i-2} \cdot p_{10}} \qquad (4\text{-}78)$$

which reduces when $p_{01} = p_{11} = p_{10} = 1/2$ to

$$\bar{n} = \frac{\sum_{i=2}^{\infty} i\ (1/2)^i}{\sum_{i=2}^{\infty} (1/2)^i} = \frac{\sum_{i=1}^{\infty} i/2^i - 1/2}{\sum_{i=1}^{\infty} i/2^i - 1/2}$$

$$\bar{n} = \ell im_{\ k\to\infty} \frac{(1+2\sum_{i=1}^{k} 1/2^{i+1} - 2(k+1)/2^{k+1}) - 1/2}{1/2} = 3 \qquad (4\text{-}79)$$

because $\sum_{i=1}^{\infty} 1/2^{i+1} = 1/2$

Now, taking into account the fact that because the profile has been over the mean line and back within three intervals means that the actual distance over the mean line is nearer two, $\bar{n}$ becomes after the end corrections have been taken away near to the value 2. Hence the dominant wavelength λ is about four times the event spacing i.e.

$$\lambda \sim 9.2\beta^* \sim 10\beta^* \qquad (4\text{-}80)$$

Consider now the power spectrum (figure 4-3). Most of the energy is contained in frequencies up to the cut-off $\omega = \frac{1}{\beta^*}$ implying an equivalent wavelength of $6\beta^*$ and longer. If the shorter wavelengths are removed by a sharp filter then the mean centre line spacing is about $5\beta^*$ (see Bendat 1963) which gives the effective wavelength of about $10\beta^*$. Hence the broadscale structure has an equivalent wavelength that is (a) about four times the distance of independence and (b) about $10\beta^*$. This gives the effective independence length to be about $2.5\beta^*$. Another way of approach is to consider the digitising of the random waveform after the high frequencies $\omega > \frac{1}{\beta*}$ have been filtered out. The necessary spacing would be about 2.5 times smaller than the cut-off wavelength which yields $\frac{2\pi\beta*}{2.5} \sim 2.5\beta*$. The value of the autocorrelation which corresponds to this is about 10% (2.3β*) which means that the independence length 2.3β* ties up with the correlation length proposed by Peklenik (1967). In this and future chapters we will most often refer to independence length rather than correlation length. This is because the words independence length convey better the physical meaning, there is no difference other than this. A statistical reason why a value of correlation between 10 and 20% should be used is that the RMS of the conditional probability (equation 4-8) becomes 95% (a well accepted confidence limit) of that of the profile.

The only difficulty with the use of low correlation values for definitions is their difficulty in measurement, however, it seems likely that they will be useful not only for the exponential but also other monotonic shapes such as the Gaussian correlation function.

4.5.2 Summary of results

The results derived from the model here presented are shown in Table 4-4. This shows the way in which the significant characteristics of a surface profile depend on the two independent parameters σ and β*. To emphasise the importance of the scale of size used in the analysis each characteristic is shown (except for the plasticity index ψ, (described in Section 4.5.3) for two scales.

TABLE 4-4

Characteristics of a random profile in terms of σ and β^*

Characteristic	Main Structure $\ell = 2.3\beta^*$	Fine Structure $\ell = 0.23\beta^*$
Mean peak height	0.82σ	0.47σ
RMS peak height	0.71σ	$0.9\ \sigma$
Ratio of peaks to ordinates	0.33	0.26
Average upward or downward slope	$0.24\ \sigma/\beta^*$	$1.66\ \sigma/\beta^*$
Mean peak curvature	$0.45\ \sigma/\beta^{*2}$	$20\ \sigma/\beta^{*2}$
Plasticity Index	$0.3(\frac{E'}{H})(\frac{\sigma}{\beta^*})$	-

The main structure is derived assuming a sampling length of 2.3β* and a fine scale structure assumes asperity dimensions one order of magnitude smaller, namely 0.23β*.

For Aachen 64-13 used in the digital analysis the shorter asperities described by 0.23β* sampling lie just within the resolution of the normal stylus. It can be seen immediately by comparison of the results that the differences between the small scale of size asperities and the main structure is considerable especially in the characteristics involving derivatives. This fact must be of consequence in contact theory.

4.5.3 Plasticity index and comparison with theory of Greenwood and Williamson

An important aim of recent studies of the topography of surfaces has been to provide an estimate of the chances that a given surface will be subject to plastic flow during contact; Blok (1952) and Halliday (1955) considered the shape of asperities which could be pressed flat without recourse to plastic deformation. It was shown that this criterion could be expressed in the form

$$\bar{m} < \underline{k}\, H/E' \qquad (4\text{-}81)$$

where H is the hardness and $E' = E/(1-\nu^2)$, E being the Young Modulus and ν the Poisson ratio; $\bar{m}$ as above is the average slope and $\underline{k}$ is a numerical factor, in the range 0.8 to 1.7 depending on the assumed shape of the asperity. Greenwood and Williamson (1966) assessed the probability of plastic deformation using their model in which asperities, each of radius R, are disposed in a Gaussian distribution of heights of standard deviation σ^*. In this model there is always a finite chance of plastic flow; however, it was shown that it depended very little upon the load but was critically dependent upon a plasticity index ψ given by

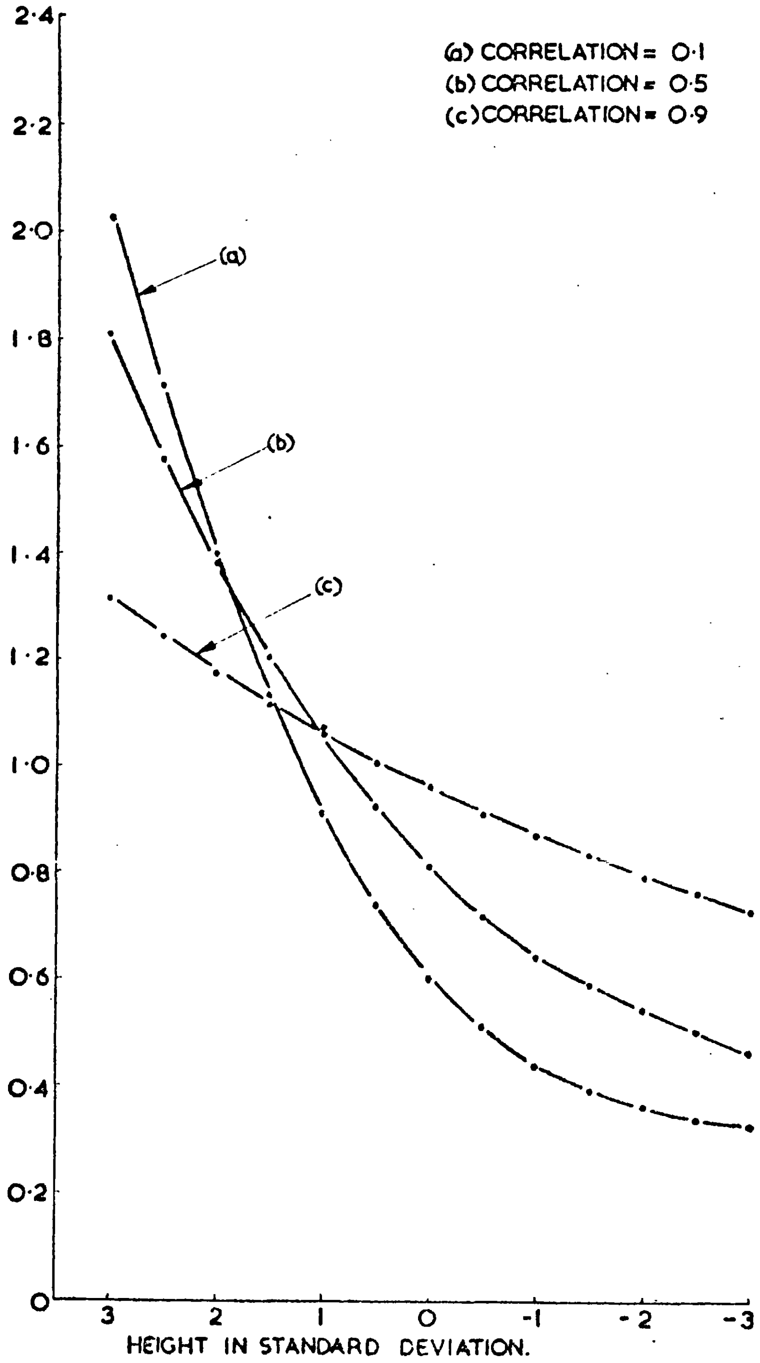
$$\psi = \left(\frac{E'}{H}\right) \left(\frac{\sigma^*}{R}\right)^{1/2} \qquad (4\text{-}82)$$

The plasticity criterion of equations (4-81) and (4-82) are similar in form. The Blok-Halliday criterion, is, however, unduly severe because it assumes complete depression of the asperities. The plasticity index of Greenwood and Williamson (1966) takes account of the fact that only the tips of asperities are normally involved in contact. The present work emphasises the simplifications which have been made in these plasticity criteria because they take no account of the existence, upon surfaces, of superposed asperities of differing scales of size, figure 4-1. The plasticity calculations of equation (4-81) and (4-82) assume that the deformation of each of the asperities is independent. Therefore the plasticity index of equation (4-82) has a significance only if it applies to the main long wavelength structure; it should then indicate the probability of plastic flow over regions associated with this scale of size. If values of R corresponding to smaller scale structure are used the arguments involved in the derivation of equation (4-82) become invalid because the deformation of adjacent asperities interact.

In deriving a value of the plasticity index ψ for Table 4-4 a value of the mean curvature of the peaks derived from

FIG 4-25

RATIO OF PEAK CURVATURE AT DIFFERENT HEIGHTS TO THAT OF MEAN VALUE



equation (4-50) has been used. The value of ψ derived in this way underestimates the probability of flow because the curvature of the peaks increases with height. Thus the highest peaks, which are those involved in contact have a smaller radius than the total peak population. Numerical integration of equation (4-46) shows the way in which the mean peak curvature varies with height. The results are shown in figure 4-25. It will be seen that for the broad-scale structure (ρ = 0.1) the upper peaks are almost three times sharper than the average. The effect of this can be seen most clearly in figure 4-23 which shows how the mean contact pressure varies with surface separation. It is clear that the agreement between the model proposed in this thesis and that of Greenwood and Williamson agree very well for the case where in our model the peak curvature is assumed to be constant. This arises from the near-Gaussian distribution of peaks obtained in our model. However, for the full model of this thesis where the curvature varies with height the curve of mean pressure against separation is different from that of Greenwood and Williamson. In fact it is about twice as high for small separations. The important feature of this new curve is that it brings out even more clearly the fact that the mean pressure at the contacts is a constant over a wide range of separations and loads.

Another feature of this model and its comparison with that of Greenwood and Williamson (1966) is worthy of comment. The Greenwood and Williamson model is specified by three parameters; σ^* the standard deviation of the peak distribution, R the radius

TABLE 4-5

Relationship between parameters of surfaces manufactured by random processes and sampled at intervals corresponding to 2.3β*

Specimen No.	Correlation (μm) Distance x 2.3	Process and Material	RMS of Peak (μm) Distribution	Average Summits per sq.cm.	Average Radius of curvature (cm)	Non-dimensional product
33-10	100	Electro-erosion-steel	9.7	1 500	0.03	0.047
34-02	30	Electro-erosion-steel	1.87	13 700	0.019	0.049
37-16	30	Electro-sinking steel	1.75	45 000	0.006	0.049
40-20	75	Electro-sinking steel	0.60	4 000	0.0136	0.034
25-18	12	Flatground-steel	0.78	450 000	0.00167	0.059
26-03	25	Flatground-steel	0.99	495 000	0.0106	0.054
27-20	15	Circum ground-iron	0.54	234 000	0.005	0.061
28-11	25	Circum ground-iron	0.86	59 000	0.012	0.050
60-04	7.5	Plunge ground-iron	0.33	495 000	0.0031	0.052
61-01	7.5	Plunge ground-steel	0.22	573 000	0.0038	0.047
64-13	15	Long ground-steel	0.35	564 000	0.0028	0.058
66-14	25	Long ground-iron	0.57	37 800	0.024	0.052

Mean value 0.050 ± 0.005

Note that for the ground specimens it is assumed when working out the number of peaks per square cm from the peaks within one track that the surface is isotropic. In fact, although this is not true for these surfaces the constant of 0.050 is still true for values along the lay where 2.3β* is much longer but by the same token the peaks go down and the radius of curvature goes up to just about equalise.

of curvature of the peaks, and η the density of asperities per unit area. In the terms of our theory the required parameters are completely defined by σ the standard deviation of the height distribution, and β^* the correlation distance. The theory of contact based on our model involves a statistical distribution of peak heights and peak curvatures. Comparing the two models σ^* is proportional to σ, R is proportional to β^{*2}/σ and η is proportional to $1/\beta^{*2}$. Hence for all random surfaces when the Greenwood and Williamson model is used the parameters should be related by the equation

$$\sigma^* \, R \, \eta = \text{constant} \tag{4-83}$$

There is some evidence (J. A. Greenwood, private communication) from the analysis of bead blasted surfaces that this relation is indeed true.

As a verification of this, Table 4-5 shows the values of this product for some random processes other than bead-blast. Each of the surfaces in this table are typical of the process.

They show that over a wide range the values do tend to be constant. For those surfaces examined the constant was 0.05 with a standard error of 0.005 which when considering the spread in individual readings taken over a surface is remarkably good.

This value of 0.05 or thereabouts could be predicted because consider the values of σ^*, R, η obtained from a random waveform, using Table 4-4 and letting the probability of an ordinate being a

TABLE 4-6

Relationship between surface parameters for surfaces generated by random processes (sampling interval 2.5 μm)

Specimen No.	RMS of peak distrib.	Average Summits per sq. cm.	Average Radius of Curvature	Non-dimensional Product
33-10	12.3	30 000	.00125	0.047
34-02	2.5	77 000	.003	0.057
37-16	2.25	35 000	.00084	0.066
40-20	0.69	720 000	.0025	0.122
25-18	1.0	1 360 000	.00055	0.074
26-03	1.25	1 440 000	.00049	0.088
27-20	0.62	1 290 000	.00083	0.066
28-11	0.90	1 270 000	.00066	0.075
60-04	0.33	1 490 000	.001	0.049
61-01	0.22	2 000 000	.001	0.045
64-13	0.4	1 800 000	.0008	0.059
66-14	0.69	1 470 000	.001	0.100

Average 0.068 ± 0.006

summit equal K, the sample interval ℓ and the correlation ρ between ordinates. Then

$$\sigma^* = \left[1 + \frac{(1-\rho)\sqrt{1+\rho}}{2\sqrt{3-\rho}\,\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}} - \frac{\pi\,(1-\rho)}{4\left(\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}\right)}\right]^{1/2} \sigma$$

$$R = \frac{1}{C^*} = \frac{2\,N\sqrt{\pi}}{(3-\rho)\sqrt{1-\rho}} \cdot \frac{\ell^2}{\sigma}$$

and $\eta = K \cdot \frac{1}{\ell^2}$ assuming unit area

here $N = \frac{1}{\pi}\tan^{-1}\sqrt{(3-\rho)/(1+\rho)}$

(4-84)

For the case where independent events are used then K = 1/5 for the first five ordinate summit model and the product becomes

$$\left[1 + \frac{(1-\rho)\sqrt{1+\rho}}{2\sqrt{3-\rho}\,\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}} - \frac{\pi\,(1-\rho)}{4\left(\tan^{-1}\sqrt{\frac{3-\rho}{1+\rho}}\right)}\right]^{1/2} \frac{2N\sqrt{\pi}}{(3-\rho)\sqrt{1-\rho}} \cdot K$$

(4-85)

$$= 0.71 \times \frac{1}{5} \times \frac{3.55}{9} = 0.05$$

Table 4-6 shows similar results obtained when sampling with an interval of 2.5 μm. This shows that even when the sampling has not been chosen for independent events the value of this product is still nearly the same being 0.068 ± 0.006 indicating that, over a wide range of size of surface roughness produced by random manufactured surfaces, and over an order of magnitude in the sampling interval, this product can be regarded as constant. Given this proviso, an estimate of K can be made for a range of interval within which the correlation is high. Remember that in the definition of the summit discussed previously in Section 4.4 the calculation of the equation for a summit for the case of high correlation was complex.

According to this the ratio of summits to ordinates for a correlation of 0.80 would be about 0.1 and not 0.004 as would result from equation (4-69). This indicates that the gross assumption made in deriving the equation, namely that only correlations to the centre ordinate were significant, is obviously not justifiable in the highly correlated case, which must be a reasonable conclusion because as $\rho_1 \to 1$ so does ρ_2 and ρ_3.

4.5.4 Comparison of theory and experiment

The comparison of theory and experiment shows that the model which has been adopted can provide a description of the geometrical features of profiles from a typical manufactured surface; the statistical distribution of surface characteristics are accurately forecast over a wide range. This range covers an order of magnitude in the linear dimensions of the asperities and more than two orders of magnitude in their curvatures. Divergences appear only at shorter wavelength and these arise, at least in part, from the resolution of the stylus.

First consider the results of figure 4-15, the exponential correlation function adopted here requires that at short sampling intervals (N→1/3) the number of peaks detected should be very nearly inversely proportional to the sampling interval. However, with the normal stylus, reducing the sampling interval from 1 to 0.5 μm increases the number of peaks by only 16% and a further reduction to 0.25 causes no detectable increase in the number of peaks. These results suggest that either the fine scale structure does not exist on the surface, or it is present and is not detected by the stylus. The results obtained with the sharp stylus show clearly that stylus resolution is a significant factor, at a sampling interval of 1 μm the replacement of the normal stylus by the sharp stylus causes an increase of 20% in the number of peaks detected. In addition, when using the sharp stylus a reduction in the sampling interval to 0.5 μm and to 0.25 μm causes increases in the number of peaks by 37 and 75%. Figure 4-14 shows that the use of the sharper stylus also results in an increase in the mean curvature of the peaks; in other words the sharp stylus reveals more detail and finer detail. That the sharp stylus can reveal detail upon surfaces which the normal stylus does not see is seen from figure 4-19 which shows a typical random surface. Also figure 4-4 shows an electron micrograph of a typical ground surface. Thus, on the question of stylus resolution we see that the effect of the finite dimension of the stylus is not likely to produce a sharp cut-off in resolution (see American Standard ASA B46.5). Nevertheless, assuming that the profile corresponds to the model of an exponential autocorrelation function, the change in the tip dimension from 2.5 to 0.25 μm produces a smaller change in the resolution than might be expected; perhaps the fine structure is present but its magnitude is smaller than is required by the theoretical model. This possibility will be further explored, and a possible explanation offered, in Chapter 7. However, it should be noted that the tip dimension may not be the only factor which determines the resolution of the stylus. Equation (4-54) shows that as the sampling interval is reduced and structures of shorter wavelength are involved the local slope of the surfaces become steeper. Hence the stylus angle might also be important in the resolution.

4.6 Conclusions

The main conclusions that can be drawn from the work in this chapter are as follows:

1. It is possible to describe those geometrical characteristics of a random surface significant in its contact properties in terms of two parameters of the profile itself, namely the RMS value and the correlation distance.

2. The scale of size is of prime importance in the determination of features likely to be important in the functioning of the surface. In particular the main structure as determined by the autocorrelation function is the most important scale of size to consider.

3. Providing that care in the definition and application of digital techniques is applied it is a powerful tool in the evaluation of surface parameters.

5. PROPERTIES CONCERNING THE MOVEMENT OF BODIES ON RANDOM SURFACES

5.1 Introduction

In Chapter 4 the evaluation of the geometrical properties of surfaces likely to be of importance in contact phenomena has been carried out using as a basic model of the surface the height distribution and the autocorrelation function. During this investigation it was found that good agreement existed between theory and practice over a wide range of sampling intervals. The major divergence between theory and experiment occurred at intervals which were of the same order of size as the stylus tip dimension and smaller. This divergence was attributed, at least in part, to the finite resolution of the stylus. One of the purposes of this chapter is to investigate, in some more detail, the characteristic of the motion of a stylus being tracked across a random surface. The approach used will not be to describe in a deterministic way the motion of the stylus, but instead, using the concept of a Markov chain, the statistical characteristics of the stylus motion will be deduced compared with the characteristics of the profile itself. It will be shown that this basic approach can also be used to deduce information about the movement of bodies much larger than a stylus and in this way it becomes possible to make useful comments about the Envelope System of surface texture assessment.

This same approach has some significance in predicting the changes that occur in surface geometry characteristics caused by rubbing surfaces under load. These questions and their significance in the running-in process will be considered in detail in Chapter 6.

FIG. 5-1

PROPERTIES OF GAP BETWEEN TWO RANDOM SURFACE

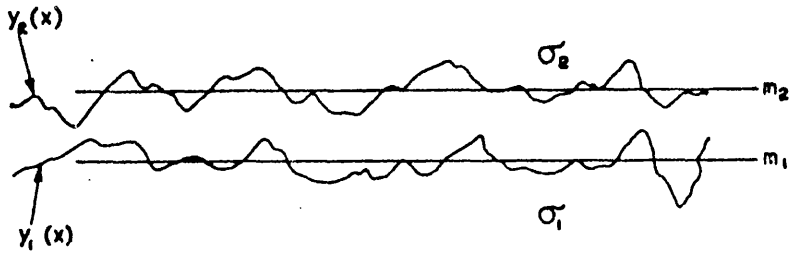


FIG. 5-2

EFFECT OF BANDWIDTH

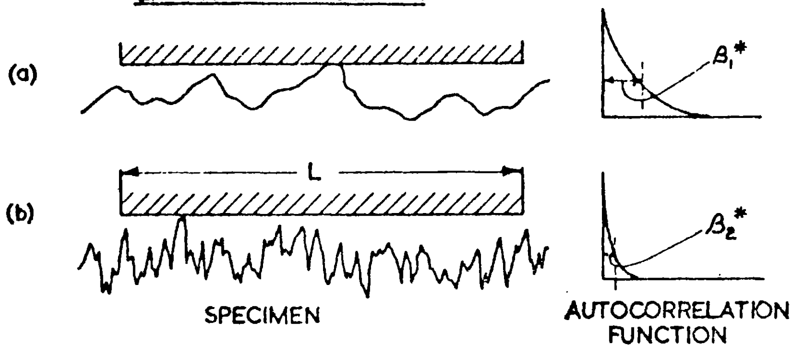
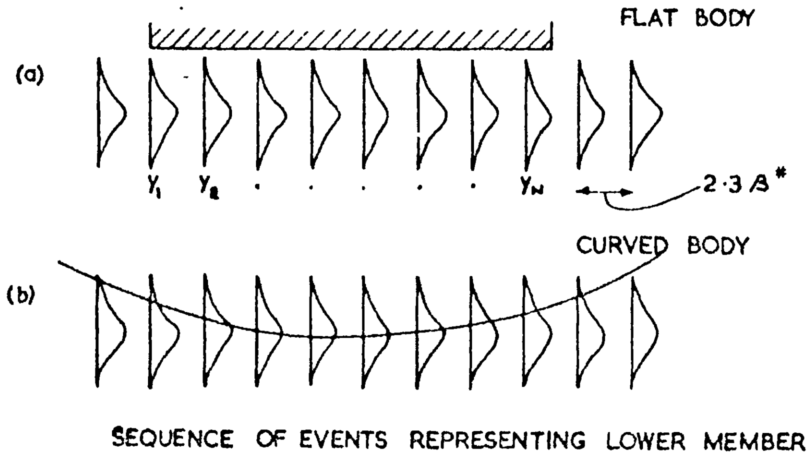


FIG. 5-3

EFFECT OF SHAPE OF UPPER MEMBER



In what follows the nomenclature given below will be used:

Subscript s - refers to stylus behaviour

Subscript e - refers to envelope behaviour

Subscript c - refers to contact in general.

Remembering also that:

No superscript refers to the profile behaviour itself

Superscript * - refers to the peak behaviour

Superscript $^{\bullet}$ - refers to the valley behaviour

Superscript ^{o} - refers to summit behaviour.

In Section 5.2, a general approach to the movement of a body over a random surface will be outlined. Following this, in Section 5.3 some specific problems will be analysed using simplified versions of this general theory.

5.2 Theory

5.2.1 Properties of gap between random surfaces

Consider figure 5-1 where $y_1(x)$ and $y_2(x)$ are random surfaces of infinite extent having probability densities of $f_1(y_1)$ and $f_2(y_2)$. If both have Gaussian ordinate height probability density functions of mean value m_1 and m_2, RMS values σ_1 and σ_2 respectively then the probability density of the gap f(g) is given by

$$f(g) = \int_{-\infty}^{\infty} f_1(y_1) \cdot f_2\,(g+y_1)\,dy_1 \qquad (5\text{-}1)$$

which yields a Gaussian distribution of mean value $m_1 - m_2$ and RMS value σg

$$\sigma_g = \sqrt{\sigma_1^2 + \sigma_2^2} \text{ so that if}$$

$$\sigma_1 = \sigma_2 = \sigma, \ \sigma_g = \sqrt{2}\ \sigma \qquad (5\text{-}2)$$

If one of the bodies is flat then $\sigma_g = \sigma$.

If the upper member (considered flat and having one degree of freedom) is moved relative to the lower then because of the infinite extent of both no vertical movement between the two would result; the top one merely resting on the highest peak of the lower. On the other hand if the upper member was of only limited length then some vertical movement would be expected because the upper member, due to its finite size, would effectively sample from the infinite population of the random surface below it. Hence statistical fluctuations in the vertical position of the flat body would result if it were moved. The magnitude of this statistical fluctuation would depend upon the relative size of the upper member to that of the bandwidth of the random signal representing the profile of the lower specimen. In two dimensions this is shown in figure 5-2 (a) and (b) in which the length of a flat of length L is compared with two random surfaces, (a) having a bandwidth - indicated here by a large correlation distance, and (b) having a short correlation distance. The flat for these purposes is restrained from tipping, a fact which is discussed later. It is clear that the statistical fluctuations in case (a) will be much greater than for case (b).

5.2.2 Effect of limited sample; independent events

Consider a random surface having a correlation distance β^*. Regions separated by about $2.3\beta^*$ can be considered to be independent of each other in the case of a first-order random surface, i.e. those having an exponential autocorrelation function. Consider therefore that the surface is comprised of a chain of such independent events. Note that these events are not necessarily peaks. The situation is shown in figure 5-3(a) in which a flat upper member whose face is parallel to the mean line of the random surface is imagined to be in contact with the surface, this in effect gives only one degree of freedom. If there are N such independent events within the sample L of the upper member, then the probability of the contact being at event y_1 between y and $y + \delta y$ is given by $f(y) \left(F\ (y)\right)^{N-1} \delta y$, where $f(y)$ is the probability density function for the independent events and $F(y) = \int_{\infty}^{y} f(y_1)\ dy_1$, is the distribution function for event 1.

A similar situation arises for y_2, y_3 and so on. Hence assuming that these individual contacts are mutually exclusive (which must be so for a system having only one degree of freedom) the probability of a <u>single</u> contact between y and $y + \delta y$ is given by $f_c(y)\delta y$ where,

$$f_c(y)\delta y \quad = \quad N\ \left(F(y)^{N-1}\right)\ f(y)\delta y \qquad (5\text{-}3)$$

Equation (5-3) holds also for surfaces having other than Gaussian height distributions and exponential autocorrelation functions.

If N is large then contact is likely to occur when y is well above the mean line in which case $f(y) \simeq 1$. Equation (5-3) then becomes (Gumbel 1959):

$$f_c(y)\delta y = N \exp \left(- (1-F(y)) \ (N-1)\right) f(y)\delta y \qquad (5\text{-}4)$$

which, for the special case of F(y) being Gaussian, immediately reveals that the probability density of a single contact at large y is more nearly exponential than that which might have been expected having a knowledge of the height density alone.

The probability density for a single contact at y, $(f_c(y))$ becomes

$$f_c(y) = N \exp \left(-(1-F(y)) \ (N-1)\right) f(y) \qquad (5\text{-}5)$$

For the particular case of a Gaussian distribution

$$f_c(y) = \frac{N}{2^{N-1}\sqrt{2\pi}} \exp(-y^2/2) \ (1 + \text{erf } y/\sqrt{2})^{N-1} \qquad (5\text{-}6)$$

From equation (5-6) it can be seen that N is the dominant factor in determining the spread of $f_c(y)$.

For an upper member which is not flat then the situation is more complicated. Consider figure 5-3(b) for instance, in which a curved body is shown being lowered onto the surface represented as a chain of independent events. In this case

$$\begin{aligned} f_c(y) &= f\,(y_1)\ F\,(y_2)\ F\,(y_3)\ \ldots\ F\,(y_N) \\ &+ f\,(y_2)\ F\,(y_1)\ F\,(y_3)\ \ldots\ F\,(y_N) \\ &\ \ \vdots \\ &\ \ \ f\,(y_N)\ F\,(y_1)\ \ldots\ldots\ldots\ F\,(y_{N-1}) \end{aligned} \tag{5-7}$$

where the individual values of y_2, y_3, etc., depend on the shape of the upper member relative to the mean line of the random surface.

Equation (5-7) reduces to

$$f_c(y) = \sum_{i=1}^{N} f\,(y_i) \prod_{j=1}^{N} \frac{F\,(y_j)}{F\,(y_i)} \qquad (5\text{-}8)$$

5.2.3 Effect of limited sample; dependent events

As before, for the sake of simplicity, we shall use a two-dimensional approach. Consider a flat upper member being lowered with one degree of freedom (y direction) on to a random profile. This precludes multiple contact unless some compliance is allowed. The events are now dependent; and, as before, designating the events, $y_1 y_2 \;\ldots\ldots y_N$ the probability of one contact in between y and y + δy at the position of event y_1 is given by

$$f_{c1}(y)\,\delta y = \int_{y}^{y+\delta y} \int_{-\infty}^{y} \int_{-\infty}^{y} \ldots\ldots \int_{-\infty}^{y} f(y_1,y_2,y_3,\ldots\ldots y_N)\; dy_N \ldots dy_1 \qquad (5\text{-}9(a))$$

Similarly for a single contact at the position of event y_2 only $f_{c2}(y)\,\delta y$ is given by

$$\int_{-\infty}^{y} \int_{y}^{y+\delta y} \int_{-\infty}^{y} \ldots\ldots \int_{-\infty}^{y} f(y_1,y_2,y_3,\ldots\ldots y_N)\; dy_N \ldots dy_2\; dy_1 \qquad (5\text{-}9(b))$$

and so on .... for a contact at the n^{th} event position

$$\int_{-\infty}^{y} \ldots\ldots \int_{y}^{y+\delta y} f(y_1,y_2,\;\ldots\ldots y_N)\; dy_N \ldots\ldots dy_2\; dy_1 \qquad (5\text{-}9(n))$$

where $f(y_1,y_2,y_3, \ldots . y_{N+1})$ is the joint probability density that y_1 has a value y_1 when y_2 has a value y_2 etc. Hence the probability that the fall of the upper member is stopped by a contact in any event position at height y is given by the sum of equations (5-9).

For the special case of a Markov process equation (5-9) can be simplified and (5-9(a)) becomes

$$= \int_{y}^{y+\delta y} f(y_1) \int_{-\infty}^{y} f(y_2/y_1) \int_{-\infty}^{y} f(y_3/y_2) \ldots . . \int_{-\infty}^{y} f(y_N/y_{N-1})\, dy_N \ldots . . dy_1$$

$$= f(y)\delta y \int_{-\infty}^{y} f(y_2/y) \int_{-\infty}^{y} f(y_3/y_2) \ldots . . \int_{-\infty}^{y} f(y_N/y_{N-1})\, dy_N \ldots . . dy_1 \qquad (5\text{-}10)$$

Equation (5-10) represents the probability that the upper member, assumed here to be a flat, will be stopped in its fall onto the lower rough member by a contact in position 1 in between height y and y + δy. In the equation the values of the other events can be anything provided that they all lie below the height y (otherwise the contact would not be at position 1). Suppose that the other events, although still being below y, had specific values, y_2 having a value in between y_a and $y_a + \delta y$, y_3 having a value in between y_b and $y_b + \delta y$ and so on up to y_N as shown in figure 5-4 then for this one particular configuration of the other events relative to y, the equation for a contact at position 1 in between y and y + δy is

$$f_{c1}(y)\delta y = \int_{y}^{y+\delta y} f(y_1) \int_{y_a}^{y_a+\delta y} f(y_2/y_1) \int_{y_b}^{y_b+\delta y} f(y_3/y_2)\cdots\cdots$$

$$\cdots\cdots \int_{y_k}^{y_k+\delta y} f(y_N/y_{N-1})\, dy_N \cdots\cdots dy_2\, dy_1$$

(5-11(a))

which reduces to

$$f_{c1}(y)\delta y = f(y)\delta y \int_{y_a}^{y_a+\delta y} f(y_2/y) \int_{y_b}^{y_b+\delta y} f(y_3/y_2)\cdots\cdots$$

$$\int_{y_k}^{y_k+\delta y} f(y_N/y_{N-1})\, dy_N \cdots\cdots dy_2$$

(5-11(b))

where k is taken to be a general alphabetic character representing the Nth event. Hence using the mean value theorem

$$f_{c1}(y)\delta y = f(y)\delta y . f(y_a/y)\delta y . f(y_b/y_a)\delta y \;\cdots\cdots f(y_k/y_{k-1})\delta y$$

(5-11(c))

Equation (5-11) can be rewritten

$$p_1 \cdot p_{a1} \cdot p_{ba} \cdot p_{cb} \cdots\cdots p_{k.k-1} \qquad (5\text{-}12)$$

where p_1 represents the probability of y_1 being in between y and y + δy; p_{a1} represents the conditional probability that y_2 is between y_a and y_a + δy (given that y_1 is in between y and y + δy); and so on. The probabilities p_{a1}, p_{ba} etc., are called transitional probabilities. Equation (5-12) represents a homogeneous Markov chain.

For the special case when the events are independent, equation (5-12) becomes

$$p_1 \cdot p_a \cdot p_b \cdot\cdot\cdot\cdot\cdot p_k \qquad (5\text{-}13)$$

In the case of a Gaussian height distribution the transitional probabilities can be written in terms of the second order normal distribution:

$$P_{a1} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \quad \exp\left[-\frac{(y_a-\rho y)^2}{2(1-\rho^2)}\right] \delta\, y \qquad (5\text{-}14)$$

Obviously equation (5-11(a)) represents only one configuration of the other events in the sequence relative to y_1. Many more configurations would have to be added in order to specify the contact situation at y_1 as fully as equation (5-10).

If we use the terminology that the value of the event at position 1 has the specific value of y_1 and y_2 has the specific value y_2 then equation (5-12) can be better written

$$p_1 \cdot p_{21} \cdot p_{32} \cdot p_{43} \cdot\cdot\cdot\cdot\cdot p_{N(N-1)} \qquad (5\text{-}15(a))$$

and the homogeneous chain probability density corresponding to equation (5-11) becomes

$$f(y) \, . \, f(y_2/y) \, . \, f(y_3/y_2) \ldots\ldots f(Y_N/Y_{N-1}) \qquad (5\text{-}15(b))$$

The concept of a first order chain can be adhered to even when the contacting event is not at the end of the chain such as y_1. Under conditions for a contact at y_m, for instance, the chain becomes

$$f(y_1/y_2) \ldots\ldots f(y_{m-1}/y_m) \, . \, f(y_m) \, . \, f(y_{m+1}/y_m) \ldots\ldots f(y_N/y_{N-1}) \qquad (5\text{-}15(c))$$

which is a property of the reversibility of Markov chains.

To see why this should be so , consider a five element chain $y_1, \ldots, y_5$ the stochastic matrix

$$M = \begin{vmatrix} E(y_1^2), & E(y_1y_2), & . & . & E(y_1y_5) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ E(y_5y_1) & . & . & . & E(y_5^2) \end{vmatrix}$$

This becomes for a Markov chain (after being normalised by σ^2)

$$\begin{vmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{vmatrix} \qquad (5\text{-}16)$$

from which the determinant is $(1-\rho^2)^4$ which gives a cofactor matrix

$$\begin{vmatrix} (1-\rho^2)^3 , & +\rho(1-\rho^2)^3 , & o , & o , & o \\ +\rho(1-\rho^2)^3, & (1-\rho^3)(1+\rho^2), & \rho(1-\rho^2)^3 , & o , & o \\ o , & \rho(1-\rho^2)^3 , & (1-\rho^2)^3(1+\rho^2), & \rho(1-\rho^2)^3 , & \\ o , & o , & \rho(1-\rho^2)^3 , & (1-\rho^2)^3(1+\rho^2), & \rho(1-\rho^2)^3 \\ o , & o , & o , & \rho(1-\rho^2)^3 , & (1-\rho^2)^3 \end{vmatrix}$$

(5-17)

The exponent in the multi-normal distribution becomes

$$\left[\frac{-1}{2(1-\rho^2)^4}\left[(1-\rho^2)^3(y_1^2+y_5^2) + (1-\rho^2)^3(1+\rho^2)(y_2^2+y_3^2+y_4^2) - 2\rho(1-\rho^2)^3(y_1y_2+y_2y_3+y_3y_4+y_4y_5)\right]\right]$$

(5-18)

This can be split up, depending on which of the ordinates $y_1 \ldots\ldots y_5$ is considered to have the contact.

Suppose for instance that y_1 makes the contact, this means that the y_1 value is fixed first, the exponent then becomes:

$$-\left[\frac{y_1^2}{2}+\frac{(y_2-\rho y_1)^2}{2(1-\rho^2)}+\frac{(y_3-\rho y_2)^2}{2(1-\rho^2)}+\frac{(y_4-\rho y_3)^2}{2(1-\rho^2)}+\frac{(y_5-\rho y_4)^2}{2(1-\rho^2)}\right]$$

(5-19)

from which the probability density becomes

$$\frac{1}{(2\pi)^{5/2}}\ \frac{1}{(1-\rho^2)}\ \exp\ (-y_1^2/2)\ \exp\ (-\ \frac{(y_2-\rho y_1)^2}{2(1-\rho^2)}\ )$$

$$\exp\ (-\ \frac{(y_3-\rho y_2)^2}{2(1-\rho^2)})\ \exp\ (-\ \frac{(y_4-\rho y_3)^2}{2(1-\rho^2)})\ \exp\ (-\ \frac{(y_5-\rho y_4)^2}{2(1-\rho^2)})$$

(5-20)

which is $= f(y_1)\ .\ f(y_2/y_1)\ .\ f(y_3/y_2)\ .\ f(y_4/y_3)\ .\ f(y_5/y_4)$

which is a true Markov chain situation for the contact at position 1. However, if y_2 had been fixed first to correspond to the contact at position 2 then the exponent becomes

$$-\left[\frac{y_2^2}{2}+\frac{(y_1-\rho y_2)^2}{2(1-\rho^2)}+\frac{(y_3-\rho y_2)^2}{2(1-\rho^2)}+\frac{(y_4-\rho y_3)^2}{2(1-\rho^2)}+\frac{(y_5-\rho y_4)^2}{2(1-\rho^2)}\right]$$

(5-21)

from which the density becomes

$$\frac{1}{(2\pi)^{5/2}} \frac{1}{(1-\rho^2)^2} \exp\left(- y_2^2/2\right) \exp\left(- \frac{(y_1-\rho y_2)^2}{2(1-\rho^2)}\right)$$

$$\exp\left(- \frac{(y_3-\rho y_2)^2}{2(1-\rho^2)}\right) \exp\left(- \frac{(y_4-\rho y_3)^2}{2(1-\rho^2)}\right) \exp\left(- \frac{(y_5-\rho y_4)^2}{2(1-\rho^2)}\right.$$

(5-22)

from which the density is seen to be

$$f(y_1/y_2) \cdot f(y_2) \cdot f(y_3/y_2) \cdot f(y_4/y_3) \cdot f(y_5/y_4) \qquad (5\text{-}23)$$

which is a Markov chain which has a change of direction at y_2. Similarly if the chain was fixed at y_m simulating the contact at y_m then $f(y_1, y_2,\ldots\ldots y_m,\ldots\ldots y_N)$

$$f(y_1/y_2)\ldots\ldots f(y_{m-1}/y_m) \cdot f(y_m) \cdot f(y_{m+1}/y_m)\ldots\ldots f(y_N/y_{N-1})$$

(5-24)

Equation (5-24) is the same as equation (5-15(c)).

In fact for N ordinates in the chain the cofactor matrix becomes:

$$\left|\begin{array}{ccc} (1-\rho^2)^{N-2}, & \rho(1-\rho^2)^{N-2} & \cdots\cdots \\ \rho(1-\rho^2)^{N-2}, & (1-\rho^2)^{N-2}(1+\rho^2) & \cdots\cdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdots\cdots\cdots & \cdots \end{array}\right| \qquad (5\text{-}25)$$

FIG. 5-4

A PARTICULAR CONFIGURATION FOR A FLAT UPPER MEMBER

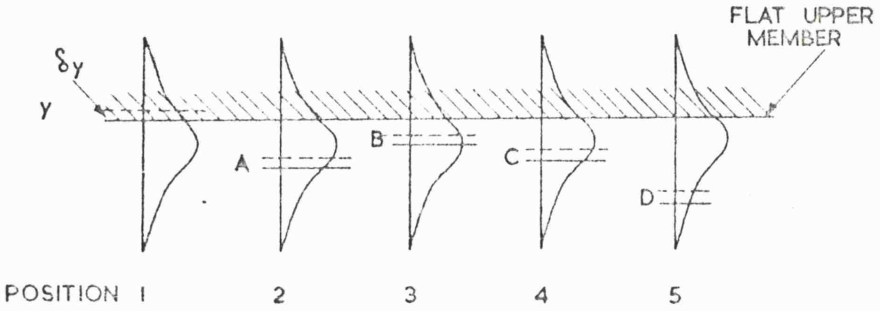
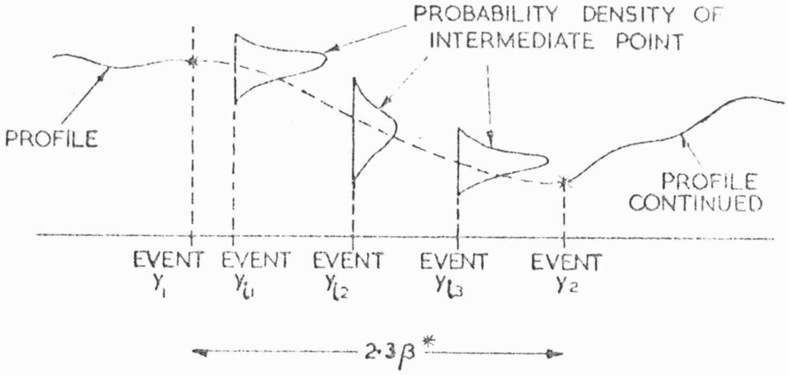


FIG. 5-5

BEHAVIOUR OF WAVEFORM IN BETWEEN INDEPENDENT EVENTS



EVENTS y_1 AND y_2 ARE INDEPENDENT

y_i IS CONDITIONAL ON BOTH

To summarise, for a surface which can be represented as a Markov type of process, that is one in which adjacent events are correlated by an exponential function, the probability of a contact between a smooth flat upper member and the profile at a height y is given by the expression

$$\int_{y}^{y+\delta y} f(y_1) \int_{-\infty}^{y} f(y_2/y_1) \ldots \int_{-\infty}^{y} f(y_N/y_{N-1})\, dy_N \ldots dy_1 \qquad (5\text{-}2\text{C}(a))$$

$$+ \int_{y}^{y+\delta y} f(y_2) \int_{-\infty}^{y} f(y_1/y_2) \int_{-\infty}^{y} f(y_3/y_2) \ldots \int_{-\infty}^{y} f(y_N/y_{N-1}) dy_N \ldots dy_1\, dy_2 \qquad (5\text{-}26(b))$$

$$\vdots$$

$$+ \int_{y}^{y+\delta y} f(y_N) \int_{-\infty}^{y} f(y_{N-1}/y_N) \int \ldots \int_{-\infty}^{y} f(y_1/y_2)\, dy_1 \ldots dy_N \qquad (5\text{-}26(n))$$

Any shaped upper member can be allowed for in equation (5-26) simply by changing the limits of integration of successive events to conform to the geometry of the upper member relative to the mean line of the surface of the lower member.

5.2.4 Behaviour of waveform between independent events

Consider an event y_i positioned somewhere in between two independent events y_1 and y_2 such that there is a correlation ρ_1 between y_1 and y_i and ρ_2 between y_i and y_2,(figure 5-5) the correlation between y_1 and y_2 being virtually zero.

Then the probability density of y_i, given specific values of y_1 and y_2, is given approximately by

$$f\ (y_i/y_1,y_2) \approx \frac{1}{\sqrt{2\pi(1-\rho_1^2)(1-\rho_2^2)}}\ \exp\left[-\frac{1}{2}\cdot\frac{(y_i-(\rho_1 y_1+\rho_2 y_2)^2}{(1-\rho_1^2)(1-\rho_2^2)}\right]$$

(5-27)

which, whenever y_i gets close either to y_1 or y_2, reduces to the familiar form

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}}\ \exp\left[-\frac{1}{2}\cdot\frac{(y_i-\rho y)^2}{(1-\rho^2)}\right]$$

Equation (5-27) represents a Gaussian distribution whose mean value depends on how close it is to y_1 and y_2 and the heights of y_1 and y_2. The standard deviation depends on the correlations between y_i and y_1,y_2. Use of this formula enables some estimates to be made of the excursions of the profile waveform between two fixed independent events.

5.3 Applications

5.3.1 Stylus resolution

The usual assumption made regarding the use of stylus instruments is that if f(y) is the height probability density of the true profile, then the probability density of the measured

FIG. 5-6

TWO EVENT MODEL FOR DETERMINING BEHAVIOUR OF STYLUS.

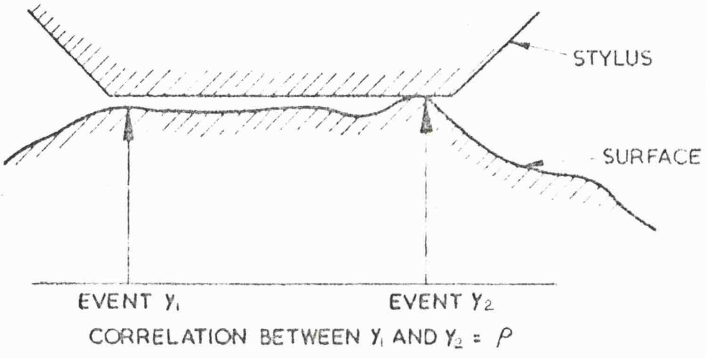
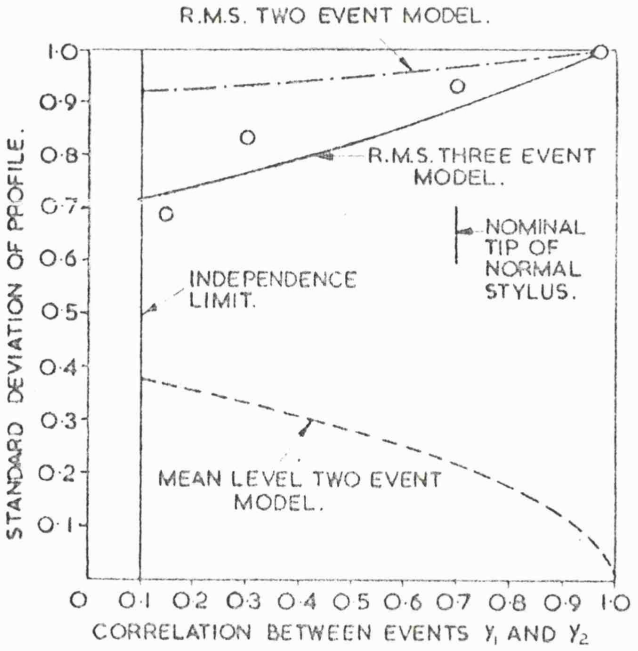


FIG. 5-7

STYLUS BEHAVIOUR TWO AND THREE EVENT MODEL

PRACTICAL POINTS MARKED O.



profile as revealed by the stylus, $f_s(y)$, is almost identical. For most practical purposes this is true enough but there may sometimes be examples where this may not be true, particularly on surfaces having a fine structure. The object of this sub-section is to use the theory given above to explore this problem of stylus resolution for random surfaces.

A stylus when tracked across a surface has one degree of freedom so that it lends itself to the model of the contact of bodies on random surfaces just given. The practical stylus has a flat tip whose dimension is usually small compared with the distance for independence on a typical random surface ($2.3\beta^*$). Indeed, it is obvious that unless this is so the stylus would be incapable of performing its intended function of exploring the surface contours; it would integrate elements of the profile. Thus the model of the profile as a chain of independent events is inappropriate for an investigation of stylus resolution. However, the model using a first order Markov chain described in Section 5.2.3 can be used.

For an approximate solution of the stylus resolution problem equation (5-26) can be used when simply two dependent events or more are considered, for instance, the stylus can be assumed to be supported by two events separated by the nominal tip dimension (figure 5-6). These two events will be highly correlated; the sharper the tip, the greater the correlation and the less the effective integration due to the stylus.

Under these conditions equation (5-26) becomes

$$f_s(y)\delta y = \int_y^{y+\delta y} \int_{-\infty}^{y} f(y_1,y_2)\, dy_1\, dy_2$$

$$+ \int_{\infty}^{y} \int_{y}^{y+\delta y} f(y_1,y_2)\, dy_1\, dy_2$$

(5-28)

which by symmetry becomes

$$= 2 \int_y^{y+\delta y} \int_{-\infty}^{y} f(y_1,y_2)\, dy_1 dy_2$$

$$= 2\, f(y)\, \delta y \int_{-\infty}^{y} f(y_2/y)\, dy_2 \qquad (5\text{-}29)$$

For a Gaussian surface the density $f_s(y)$ becomes

$$f_s(y) = \frac{\exp\,(-y^2/2)}{\sqrt{2\pi}} \left(1 + \operatorname{erf}\,(y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right) \qquad (5\text{-}30)$$

which, because

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\,(-y^2/2) \left(1 + \operatorname{erf}\,(y/\sqrt{2}\sqrt{\tfrac{1-\rho}{1+\rho}})\right) dy = 1$$

is a true probability density.

Equation (5-30) represents the probability density that a stylus having a flat tip will fall to a height y (where y is measured from the mean line of the surface).

Taking the first moment of equation (5-30) gives the mean height of the stylus above the mean line $\bar{y}_s$ i.e.

$$\bar{y}_s = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \; y \;\; \exp\,(-y^2/2)$$

$$\left(1 + \operatorname{erf}\,(y/\sqrt{2}\,\sqrt{\tfrac{1-\rho}{1+\rho}})\right) dy$$

$$\bar{y}_s = \sqrt{\frac{1-\rho}{2\;\pi}} \qquad (5\text{-}31)$$

where this is a ncn-dimensional value being expressed as a fraction of the RMS value of the true profile.

Similarly the RMS value of the measured profile σ_s, again expressed as a fraction of the RMS of the real profile is given by the second central moment.

$$\sigma_s^2 = \int_{-\infty}^{\infty} y^2 \; f_s(y)dy - \left(\int_{-\infty}^{\infty} y \; f_s(y)dy \right)^2$$

$$= 1 - (\frac{1-\sigma}{2\;\pi})$$

from which

$$\sigma_s = \sqrt{1 - (\frac{1-\rho}{2\;\pi})} \qquad (5\text{-}32)$$

Equations (5-31) and (5-32) are plotted in figure 5-7. They show that at conditions where $\rho \sim 1$ when the two events are close and the stylus tip is sharp the mean line of the measured profile approaches that of the true profile i.e. $\bar{y}_s \sim 0$. Also under these conditions the RMS of the measured profile approaches that of the true profile i.e. $\sigma_s \sim 1$. Both of these limits are physically sensible. Consider as an example the case when a typical stylus having a tip dimension of 2.5 μm is tracked across Aachen 64-13 whose independence (correlation) length is 15 μm. For this situation $\rho \sim 0.7$ and according to the formula (5-32) the RMS value as measured would be about 2% less than that of the true profile. Measured values for this stylus and a few others, specially manufactured for this exercise, are shown marked on figure 5-7. While it is possible to get an idea of the RMS reduction brought about by having a stylus not infinitely sharp, it is not so easy in the case of estimating the mean level shift. To estimate the RMS value a large number of tracks were made in a well-defined region on the surface for a number of styli; one being really sharp (say, a tip dimension of 0.25 μm). The mean RMS value for readings taken with this stylus was taken to be unity. RMS values for similar readings taken with the other styli were then compared to this. Two points are worth noting. First, it was possible to use a sharp stylus on the Talysurf in this case because the specimen was not isotropic (being ground) and consequently a large tip dimension parallel to the lay was possible (7 μm) which gives the necessary strength. Second, it is not possible to do a similar exercise for the mean height because of the difficulty in relocating the pick-up in the same place after changing the stylus.

Obviously figure 5-7 only represents an approximation to the true stylus behaviour both at the very high and low correlation regions of the graph. At very high correlations it has already been pointed out in Chapter 4 that it is not only the tip dimension that influences the stylus resolution, the stylus angle can also have an effect. Further, this effect is even more marked as the RMS value of the surface increases without an increase in β^*. At the low correlation end of the graph, as the stylus gets bigger, the two events approach a separation where they can be considered to be independent; then a new model must be used incorporating a number of independent events, i.e. use would then be made of equation (5-6).

As a close approximation to stylus behaviour more than two events supporting the stylus can be used in which case the analytical solution becomes more difficult. However, use can be made of a result derived in Chapter 4 for the definition of the RMS value of the peak distribution using three-point analysis. This corresponds closely to the three-event support for the stylus. Hence under these conditions.

$$\sigma_s \sim \left[1 + \frac{6}{\pi}\,\frac{(1-\rho)\sqrt{1+\rho}}{(4-\rho)\sqrt{3-\rho}} - \frac{36}{\pi}\,\frac{(1-\rho)}{(4-\rho)^2}\right]^{1/2} \qquad (5\text{-}33)$$

Notice that when the three-event model is used a stylus shape which is not necessarily flat but quadratic could be catered for. The result for three horizontal points is shown plotted on figure 5-7. The agreement between the practical results and the theory now appears to be closer.

The conclusion to be reached from the work in this section is that the stylus commonly used in surface texture instruments of tip dimension 2.5 μm only affects the assessment of the texture by a small amount.

5.3.2 Envelope system of assessing surface texture

In Section 5.3.1 an example has been given in which it is useful to regard the surface profile as a first order Markov chain; the events being correlated with each other. In this section an example will be given of how the surface profile can be usefully taken to be a succession of independent events for contact phenomena. Beckmann (1957 and 1959) has attempted to apply concepts similar to these in the problem of reflection from a single flat rough surface.

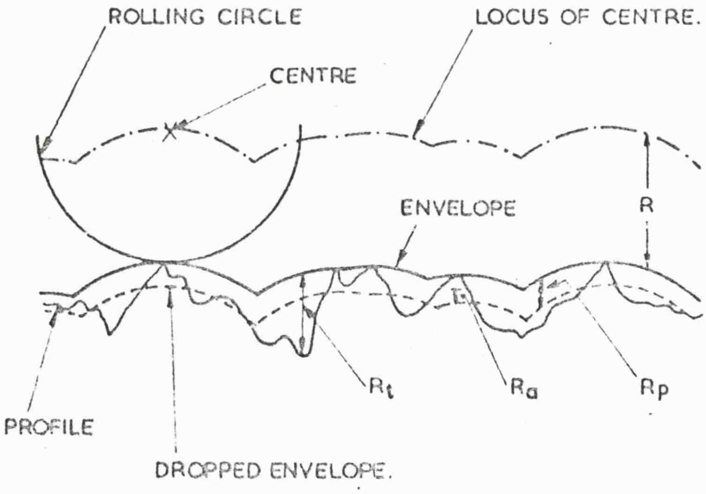
The Envelope or E-System of assessing surface texture was devised a number of years ago in Germany (von Weingraber 1957). It is a system devised to remove the unwanted long wavelengths from a profile graph prior to assessment. It differs from the now internationally adopted Mean Line or M-System (B.S. 1134) in that the reference line that is constructed from which the texture is evaluated is not a mean line but an envelope or crest line.

In the E-System a graphical construction is made on the profile graph which simulates the rolling of a circle of fixed radius across the top of the profile waveform. The locus of the lowest point on the ball or circle is then taken as the reference from

FIG.5-8

ENVELOPE SYSTEM OF ASSESSING SURFACE TEXTURE.

R IS THE RADIUS OF THE ROLLING CIRCLE
Rp IS THE AMOUNT BY WHICH THE ENVELOPE HAS TO BE DROPPED IN ORDER TO MAKE IT A MEAN LINE.
Ra IS THE AVERAGE HEIGHT OF THE PROFILE FROM THE DROPPED ENVELOPE.
Rt IS THE MAXIMUM DEPTH OF THE PROFILE FROM THE ENVELOPE.



which the texture is measured, figure 5-8.

The use of the envelope just described as a datum from which measurements are made effectively filters out the long wavelengths; how long these wavelengths are is determined by the radius of the circle. Obviously, as in the M-System where a number of cut-offs for the filters have to be used, more than one radius has to be used in the E-System. The usual convention is to use two radii, one at 50 mm and the other at 3.2 mm; the former to isolate the waviness and the latter the roughness. As previously stated the R_p value is the difference between the mean height of the envelope and the mean of the profile. The R_t value is the maximum difference between the envelope and the deepest valley. The R_a value (which should have a suffix E on it) is the same, virtually, as for the M-System only the mean line now is the envelope line which has been dropped from the crests in such a way that it splits the profile into two parts; the area between the profile and envelope being the same above the envelope as below it. (figure 5-8).

It is of considerable importance to determine the properties of the Envelope System for random waveforms rather than deterministic (Reason 1962) so that a better comparison can be made between it and the M-System.

In order to get an insight into the way in which the R_p value changes with radius, and also incidentally how the mean separation between bodies changes with shape, consider how the mean value of the envelope relative to that of the profile can be estimated.

Consider, first, the flat body case, using equation (5-3) the mean height of the envelope $\bar{y}_e$ is given by

$$\bar{y}_e = \int_{-\infty}^{\infty} y\ N\ F(y)^{N-1}\ f(y)\ dy \qquad (5\text{-}34)$$

This is difficult to evaluate for N having other than very small values such as might be the case for a very blunt stylus. This is still true for our specific model where it assumed that the ordinate height probability density function is Gaussian and equation (5-34) reduces to

$$\bar{y}_e = \int_{-\infty}^{\infty} y \frac{N}{2^{N-1}} \left[1 + \operatorname{erf}\ (y/\sqrt{2})\right]^{N-1} \frac{\exp(-y^2/2)}{\sqrt{2\pi}} \qquad (5\text{-}35)$$

Under normal conditions F(y) in equation (5-34) varies slowly for points far removed from the origin, i.e. large y. The mean value $\bar{y}_e$ is then very close to both the median and the mode of the distribution. To attempt to work out the median rather than the mean is not a solution of this difficult problem. The median is the value y_{med} such that

$$\int_{-\infty}^{y_{med}} N\ F(y)^{N-1}\ f(y)\ dy = 1/2 \qquad (5\text{-}36)$$

which seems equally difficult to evaluate. However, the modal value (that value of y for which the probability density is a maximum) is much simpler to find; it can be achieved by differentiating y_e.

This yields the condition for the modal value y_{mod}

$$(N-1)\left[f(y_{mod})\right]^2 + F(y_{mod})\ f'(y_{mod}) = 0 \tag{5-37}$$

which for equation (5-6) becomes

$$y_{mod}\ \exp\ (y^2_{mod}/2)\ \left[1 + \operatorname{erf}\ (\frac{y_{mod}}{\sqrt{2}})\right] = (N-1)\sqrt{\frac{2}{\pi}} \tag{5-38}$$

To see how y_{mod} is influenced by N an additional assumption can be made, namely that in most cases $1 < y_{mod} < 4$. This means that $(1 + \operatorname{erf}\ (y_{mod}/\sqrt{2}))$ can be considered virtually constant $\sim \sqrt{2}$ within a few percent. Hence

$$y_{mod}\ \exp\ (y^2_{mod}/2) \sim \frac{N-1}{\sqrt{\pi}} \tag{5-39}$$

Taking logarithms yields

$$2\ \ln\ y_{mod} + (y_{mod})^2 = 2\ \ln\ (N-1) - 2\ \ln\ \sqrt{\pi} \tag{5-40}$$

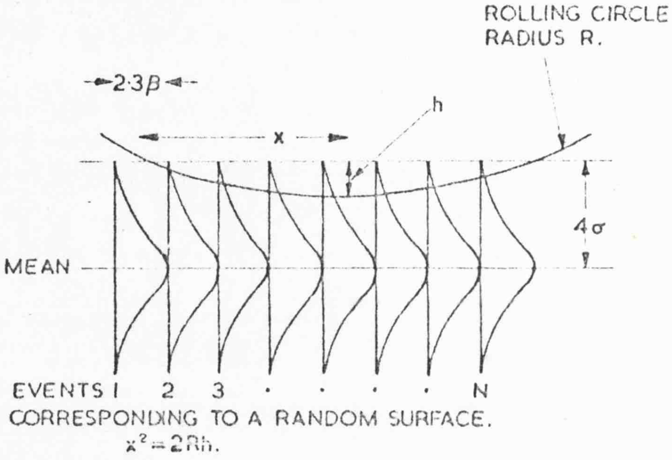
Now since $\ln y^2 << y^2-1$ if $y>0$ it can be seen that to a good approximation

$$y^2_{mod} \simeq \ln N^2 \tag{5-41}$$

To extend this to the situation found in the E-System where the upper body is circular we need to take account of the shape. To do this the simple spherometer formula can be used;

FIG. 5-9

COMPUTATION OF ENVELOPE PROPERTIES.



i.e. $x^2 = 2Rh$ where x is the semichord length, R is the radius and h is the distance from the chord to the parallel tangent at the circumference.

Consider now a simplified treatment based on the spherometer formula. We require an effective value of N, the number of independent events involved when a circle of radius R rolls over the surface profile at a depth h from the crests. In this simplification the curved member will be considered as the equivalent of a flat member of length 2x where $x^2 = 2Rh$. Also the number of independent events N in the length 2x is $2x/2.3\beta^*$ where $2.3\beta^*$ is the independence (correlation) length for an exponential correlation function. This is shown in figure 5-9.

From these assumptions an estimation of the behaviour of $\overline{y}_e$ can be made, thus

$$N^2 = \frac{4x^2}{(2.3\beta^*)^2} = \frac{8Rh}{(2.3\beta^*)^2}$$

$$= 8 \cdot \frac{R}{2.3\beta^*} \cdot \frac{h}{2.3\beta^*}$$

$$= \left(\frac{R}{2.3\beta^*}\right) \cdot K_e$$

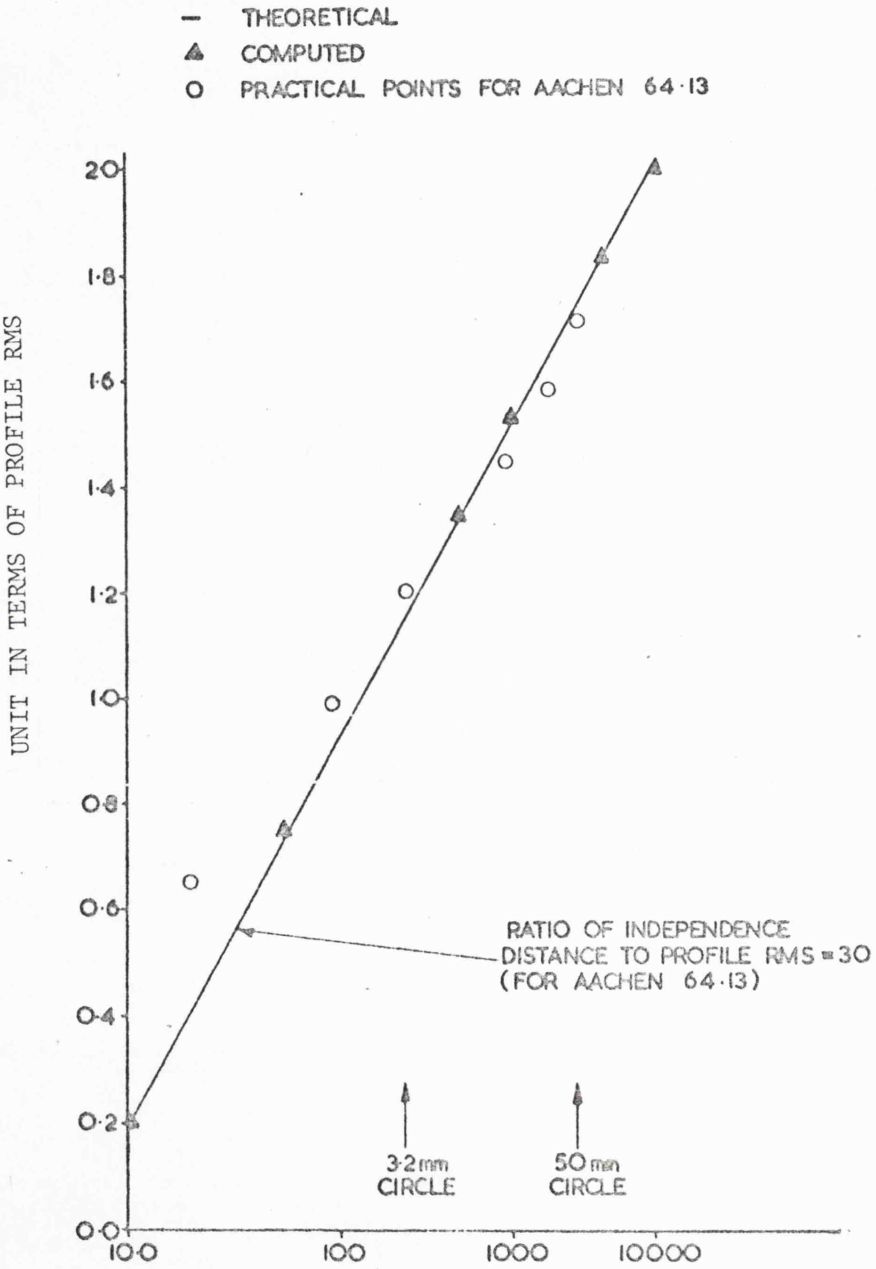
Hence

$$y^2_{mod} \sim \overline{y}^2_e \sim \ln \left(\frac{R}{2.3\beta^*}\right) + \ln(K_e) \qquad (5\text{-}42)$$

In practice K_e has a value less than 0.3 whereas $(R/2.3\beta^*)$ has a value typically of 500 to 1000. Thus in equation (5-42), the

FIG. 5 - 10

MEAN SEPARATION OF ENVELOPE AND PROFILE (Rp)



RATIO OF RADIUS TO INDEPENDENCE DISTANCE
(PLOTTED ON A SQUARE-ROOT-LOGARITHMIC SCALE)

term of greatest significance is ln $(R/2.3\beta^*)$. Thus a rough estimate of the way in which $\bar{y}_e$ (or R_p) changes with the radius is given by

$$\bar{y}_e = R_p \; \alpha \; \sqrt{\ln(R/2.3\beta^*)} \qquad (5\div43)$$

To test this statement equation (5-8) was solved numerically taking the curvature into account for a wide range of values of R using a program called REST (Appendix 2). Some of the results are shown in figure 5-10, the mean shift between the random Gaussian profile and envelope as simulated by the computer is plotted against $\sqrt{\ln(R/2.3\beta^*}$. It can be seen that, over the three decades of radius plotted in figure 5.10, a linear relationship does appear to hold for a random surface and therefore the approximations which were made in equations (5-41) and (5-43) seem to be justified.

Figure 5-10 also shows some results obtained from a real surface. Three tracks on Aachen 64-13 were data logged and the process of rolling different circles across the profile graphs carried out by the computer using program ROLL (Appendix 2). The parameters thus generated were outputed to allow a comparison with the theoretical predictions mentioned above. For Aachen 64-13 all the relevant parameters such as β^* and σ were known to enable the practical points of R_p to be plotted for different values of radius. It can be seen that there is a fair agreement between theory and practice especially at large values of R; large divergences occur for small R.

Given the values of σ and β^* for Aachen 64-13 then according to the approximate theory plotted as a continuous line in figure 5-10 the ratio of the R_p values obtained on Aachen 64-13 using the two standard radii 50 mm, and 3.2 mm should be 1.53. The practical ratio worked out as 1.45, a divergence of about six percent. The R_p values themselves were 1.15 and 1.75 for the theoretical results of the 3.2 and 50 mm circles respectively and 1.19 and 1.73 in the practical simulation.

In conclusion it appears that the theory given above governing the behaviour of the Envelope System does have some basis in fact. Divergences are to be expected to some extent because of the presence of abnormally high peaks or lack of uniformity in the three tracks taken. In order to compare the envelope behaviour as used in the E-System with the equivalent M-System results, it would be necessary to examine surface parameters such as R_p and R_a not only relative to the envelope but also relative to the standard wavefilter mean line and centre line.

5.3.3 Other bodies

The movement of the upper body when both of the bodies are rough is an additional problem of significance. Suppose that the two bodies have different bandwidths, i.e. different values of β^*, it is always possible to get the geometry of the gap simply by subtracting the one waveform from the other. Under these circumstances the original situation can be replaced by one in which a flat surface is contacting a random surface - whose geometry is that of the gap between the two original surfaces. Similar statements also apply to other derived properties of the gap. Thus the curvature of the gap between random surfaces will be the sum of the curvatures of the two original random surfaces.

Unfortunately it is not straightforward to work out parameters of the movement or positioning of one of these rough bodies on another when, say, the upper one has other than a flat general shape, or is of limited length. A number of researchers, namely Iwaki and Mori (1958), Tsukizoe and Hisakado (1965), Kimura (1966) have investigated problems similar to these but under loaded conditions, and all have assumed that in the region of contact the mean levels of the two surfaces are parallel. In the case of one rounded member the elastic conformity due to the loading is equivalent to an assumption of parallelism. Under these circumstances the number of contacts within the given contact length are either worked out using the level-crossing theorems due to Rice (1944 and 1945) or are assumed from an estimated number of asperities per unit length. However, in the case that we have considered where no load is applied, the shape of the upper member cannot be ignored and this is one of the few examples where the no load condition is more difficult to deal with geometrically than the highly loaded situation.

5.4 Discussions and Conclusions

In this chapter it has been indicated that, by condensing a profile into a succession of events, it has been possible to tackle problems, even if in a simplified way, that otherwise would have been too difficult. A special feature has been that the effect of limited size of specimen and general shape of specimen have been considered with reference to the behaviour of one body moving or positioned on another.

In particular in the case when the events making up the surface have been considered to be correlated it has been possible to investigate the behaviour of styli moving across random surfaces - a problem having considerable significance in the field of surface metrology. Also, when the events have been considered to be independent it has been possible to investigate some of the properties of the Envelope System of surface texture assessment - again an important problem in surface metrology. In other fields like those of thermal and electrical contact these concepts may be useful, especially where problems of average clearance is important (Howells and Probert 1968).

This chapter has been concerned with contact without deformation or wear. Clearly when deformation and wear occur the overall problem is usually much more difficult. Therefore, in the next chapter we will start by considering experimental results which occur when bodies are rubbed under load.

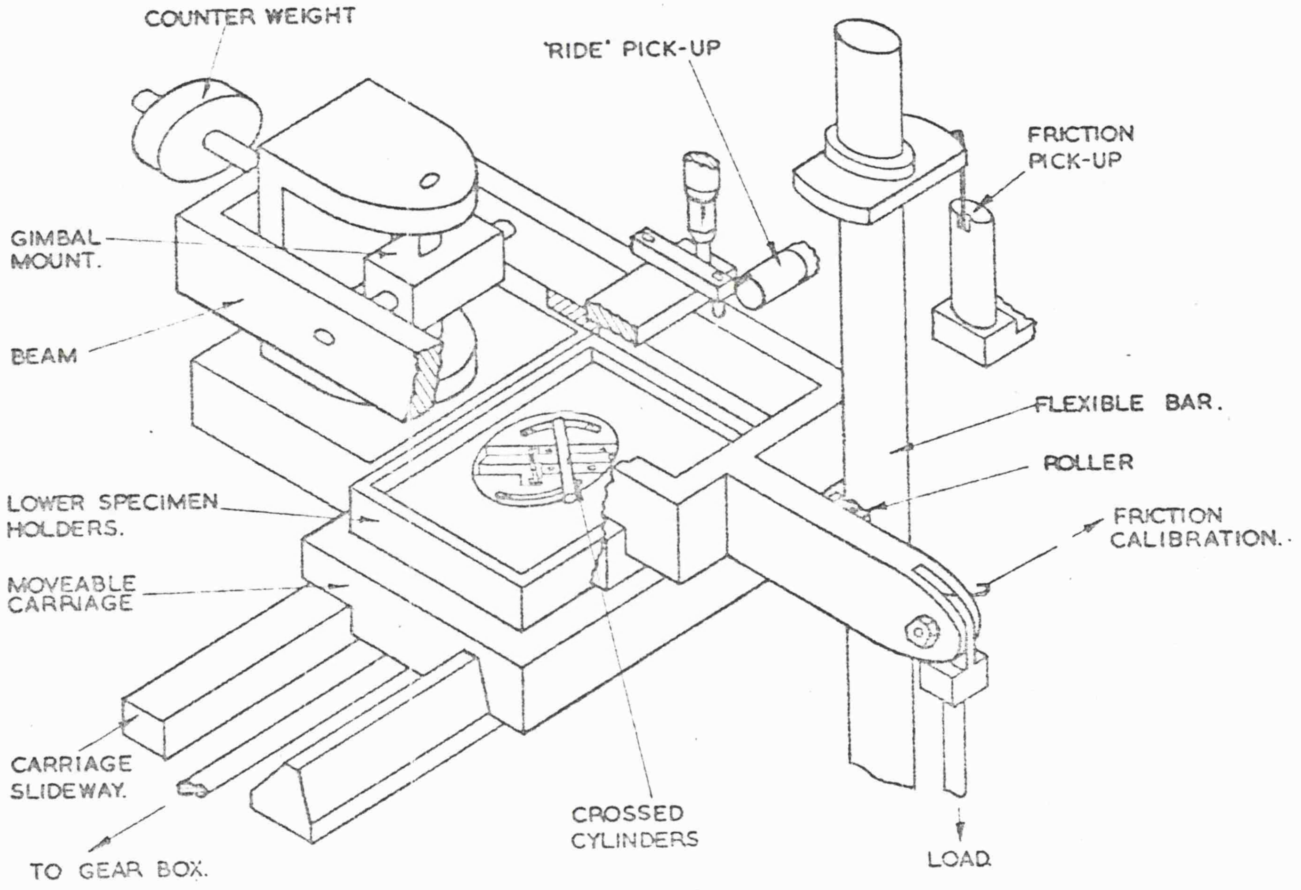
6. SOME EXPERIMENTS IN WEAR AND FRICTION

6.1 Introduction

In Chapter 5 the way in which a body contacts with a random surface has been examined, both for the case where the dimension of the body is small compared with the correlation length of the surface and also for the case where it is much longer. No attempt, however, was made to investigate the contact under load; the physical situation occurring in wear and friction. These problems will be investigated in this chapter. The approach adopted will be based upon simple straightforward experiments in which the major measurements will be of the friction between the bodies and the changes which occur in the surface topography as a result of rubbing. The extent to which these measurements can be interpreted in terms of the stochastic concepts of earlier chapters will be explored.

In investigating the changes in topography due to rubbing we shall use the term "wear" even though it is not established that such changes in surface geometry arise from a complete removal of material as loose debris. These changes in surface topography are, of course, part of the process which is generally referred to as "running-in". In order to be able to monitor accurately those changes it is necessary to have a technique in which the geometry of a particular track in the specimen can be examined during the wear process. Such a technique will now be described.

FIG 6-1 CROSSED CYLINDERS MACHINE



6.2 Experimental apparatus and techniques

6.2.1 General Principles

Ideally, an instrument to measure the surface geometry of a particular track during running-in would be an integral part of the machine in which the experiment was taking place, (in this case the experimental rig was the Archard crossed-cylinders friction machine, figure 6-1). However, because the integration of the Talysurf and the crossed-cylinders machine proved to be impractical, a relocation technique had to be devised which enabled a specimen to be measured in the Talysurf and rubbed in the crossed-cylinders machine in such a way that during the running of an experiment not only was the same track always measured on the Talysurf but exactly the same part of the specimen was rubbed in the crossed-cylinders machine. This means that two relocation schemes are needed, one for Talysurf and one for the crossed-cylinders machine, together with a method of relating the measured track on the one to the rubbed track on the other.

Before explaining these techniques a short description of the crossed-cylinders machine will be given together with the functional features that, after modification, it was able to measure.

6.2.2 The crossed-cylinders friction machine

A description of the machine and its advantages have been given by Archard (1958). This consists essentially of a movable carriage upon which a specimen can be mounted; over and at right

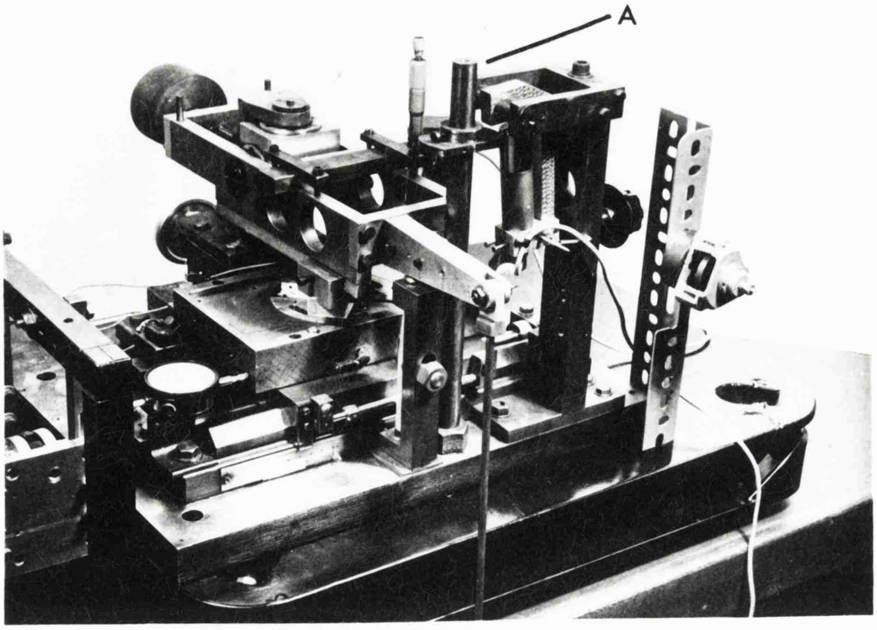


Figure 6-2. Crossed Cylinders Machine showing flexible bar (A).

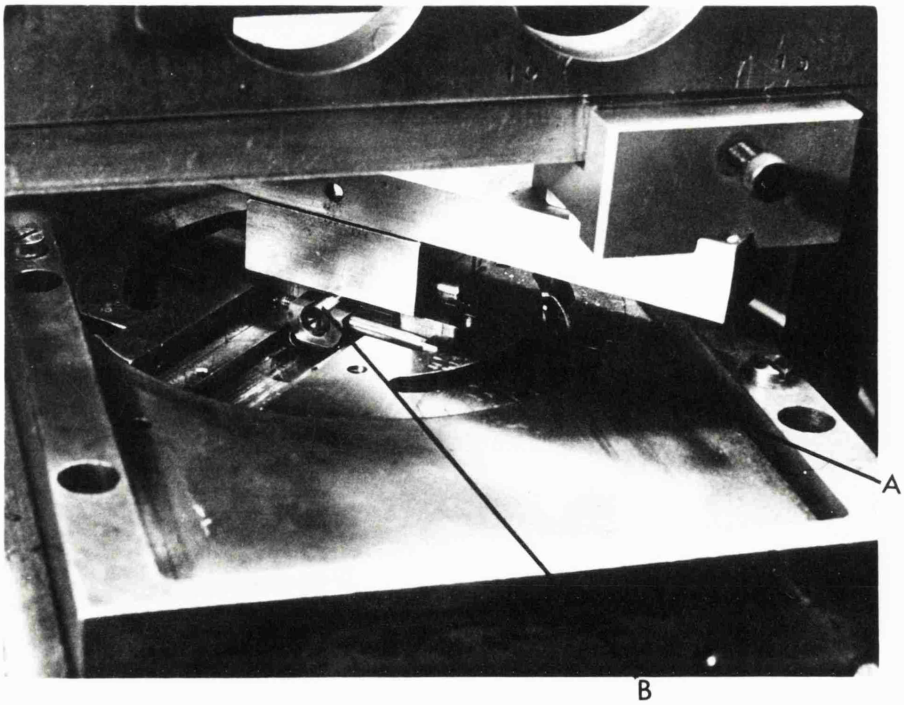


Figure 6-3. Crossed Cylinders Machine.

(A) Upper specimen

(B) Lower specimen

angles to the carriage a beam is mounted on gimbals. Another specimen is hung under the beam so that it can make contact with the upper surface of the specimen on the carriage. In this arrangement the lower cylindrical specimen is in a horizontal plane with its axis at 45° to the direction of traverse of the carriage while the upper member, also in a horizontal plane is fixed at 45° to the beam in such a way that it makes a right angle with the lower specimen. (Figure 6-3). Thus the area of contact between the two specimens is a circular region into which any part of either specimen only enters once during a sliding operation, figure 6-14(c). This has considerable advantages in experiments of this kind because it helps to prevent avalanche processes, which are sometimes exhibited in wear experiments.

Various loads can be applied to the specimens by loading the free end of the beam. Also, lubrication conditions can be altered from dry to boundary because the lower specimen is held in a holder, figure 6-4, which has a lip surround which is higher than that of the contact point between the specimens. By filling the holder with oil the conditions can easily be changed from dry to boundary lubrication. The holder also has to incorporate other features described later.

An additional feature of the machine is a friction measuring device in which the frictional force between the two specimens, when in contact and moving relative to each other, causes a deflection of a flexible bar A. (Figure 6-2). This deflection is monitored by a Talysurf pick-up, the resulting signal from which is amplified by another Talysurf amplifier unit and then recorded. There is, in addition, one more characteristic which can be measured; this is the vertical movement of the beam caused by the one specimen running across the top of the other. Obviously if both specimens were smooth and straight no such movement would arise. Here, as for the friction measurement, a Talysurf pick-up and amplifier are used. An alternative device used sometimes to measure the vertical movement, the "ride", is the Talymin Side Acting Gauge.

One practical feature that has to be incorporated into the friction measuring system is a high-cut filter, usually in our case a digital filter, which is sometimes necessary to remove the oscillations produced by resonances in the bar. Although generally this filter is not required, in the case of light loads the frictional force is so small that a thin bar is required in order to allow the frictional force to produce a sufficient deflection. This in turn can have the effect of lowering the resonant frequency of the friction measuring system into the pass-band of the friction signal which could, if undetected, give rise to false friction readings.

The output from the friction and the "ride" systems can be linked to the data logger; the output signals from both being compatible with that emerging from the normal Talysurf.

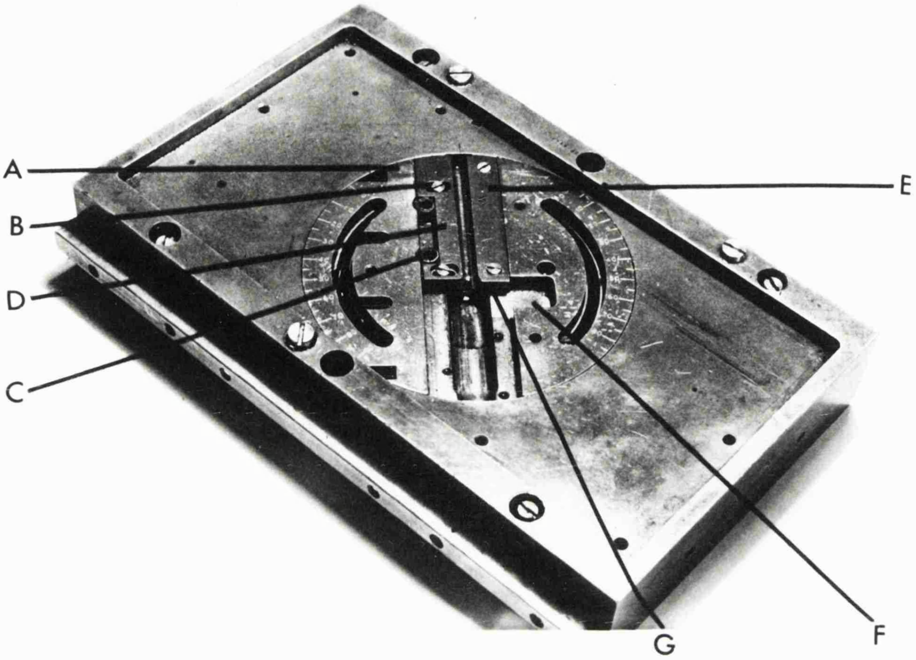


Figure 6-4. Lower specimen holder.

(A) Rotatable table.

(B) Screws to hold jaws down.

(C) Eccentric heads on screws.

(D) Movable jaw.

(E) Fixed jaw.

(F) Flat to locate specimen rotationally.

(G) Flat to locate specimen axially.

6.2.3 Relocation profilometry apparatus

For any given track along the lower specimen easy and direct comparison of any of the existing surface finish parameters is possible because the information once stored digitally can be re-examined time after time with different parameters in mind, using exactly the same data. Hence for any given profile of the texture any comparison of parameters is likely to be very accurate, depending on the quality of the numerical analysis techniques that have been adopted. However, useful comparison depends on the quality of the information in the first place.

In the basic experiment the surface finish of the lower specimen is measured on the Talysurf. It is then put in the holder (figure 6-4) on the carriage. The specimen is rubbed once, the friction and ride graphs usually being taken at the same time. The specimen is then removed from the holder and re-measured on the Talysurf. After this the specimen is repositioned in the holder and the process repeated perhaps with an increased load. The specimen is re-measured after one, two, four rubs etc. From the graphs or digital data the relevant parameters can be measured throughout the whole of the running-in process.

The success of this type of experiment relies not only on the wear occurring in the same place after repositioning but also the measurement taking place along the same track. To achieve these conditions some kinematic features had to be incorporated into both the friction machine and the Talysurf vee block

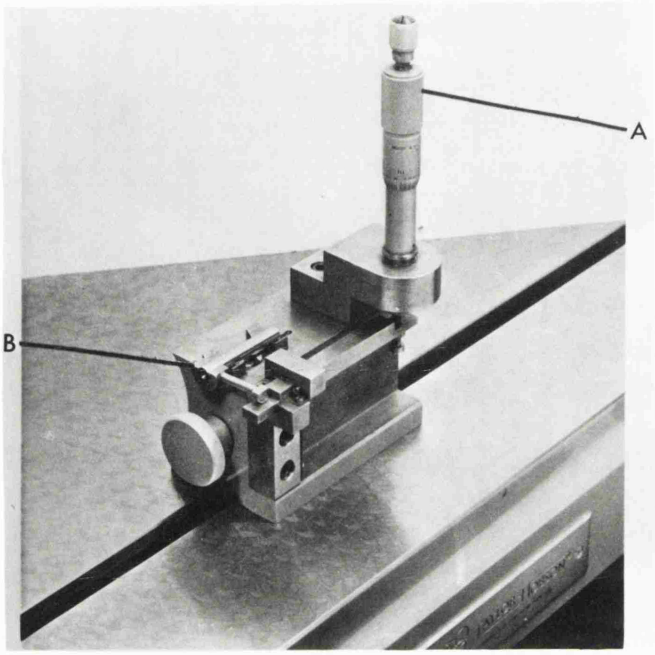
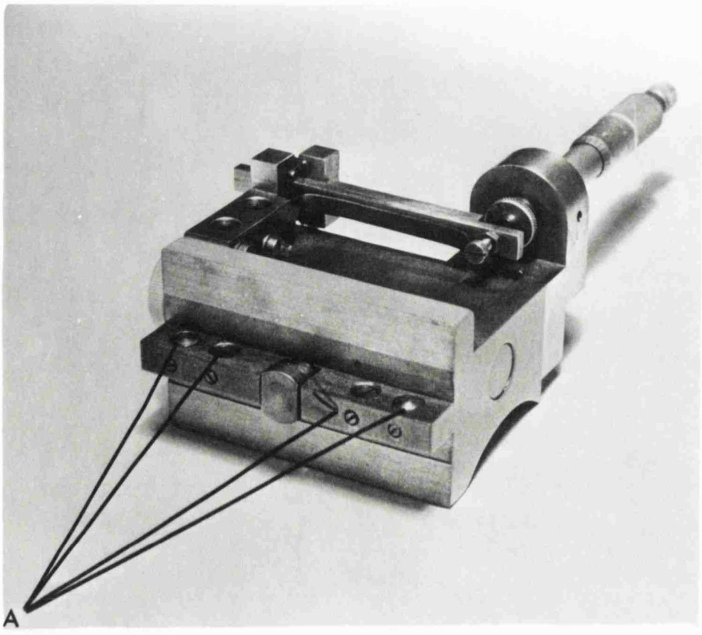


Figure 6-5.
Modified Vee Block.
(A) Micrometer.
(B) Specimen Collar.

Figure 6-6. Modified Vee Block
(A) Positioning screws in base.



arrangement. In order to facilitate this a three-quarter collar was put on the end of the lower-removable-specimen. This is marked with the letter B in figure 6-5. Fixed to the collar is a hardened stud.

In the case of the Talysurf relocation problem the necessary modifications had principally to be made to the vee block, figures 6-5 and 6-6. This involved designing a system giving six constraints. To do this the vee itself was cut away to provide effectively two sets of knife edges; contact of the specimen against these edges was achieved by means of a magnet inserted into the base of the block. Axial movement of the specimen was inhibited by locating the collar against a milled flat on the end of the vee block. The stop inhibiting the rotational movement was made adjustable tc allow any region of the wear track to be investigated. The stop was attached to the side of the vee block onto which the specimen collar stud located and was in the form of a strip of metal, hinged, via crossed ligaments, to a fine micrometer adjustment at the rear of the block, figures 6-5 and 6-6 Movement of the micrometer produced a small rotational adjustment of the specimen. This proved useful in the investigation of the metal flow within the wear region.

The same principle was employed in the specimen holder of the crossed-cylinders machine, figure 6-4, flats F and G being used to locate the specimen rotationally and axially. A pair of jaws, one movable, held the specimen. Knife edges were not possible in this arrangement because of the load sometimes imposed on the specimen.

Eccentric screw heads on the movable jaw allowed easy release of the specimen from the holder.

These features ensure accurate repositioning of the specimen after each wear rub, and measurement. It was equally important that the Talysurf should have measured that part of the wear track which was of most interest. Before this could be achieved, however, the measured track on the Talysurf had to be set up correctly. This in itself involved making the track of the pick-up parallel to the tee slot in the Talysurf base, then ensuring that the vee block axis was parallel to the tee slot and finally that the track of the Talysurf pick-up lay in the same vertical plane as the axis of the vee block. The first part was accomplished by using a clock gauge which was positioned to measure between the edges of the tee slot and the gearbox column base. Positioning of the vee block in the tee slot was achieved by means of screws set into the base of the vee block, figure 6-6. The two outer screws were used to adjust the direction of the vee block relative to the tee slot whilst the two inner ones were used to take up excessive slack.

A straight, smooth specimen of the same diameter as the test specimens (62 mm) but about one third of a metre long was set in the vee block and the screws adjusted until the long specimen was the same distance from the second tee slot all along its length. The vertical position of the gearbox was also adjusted at this stage.

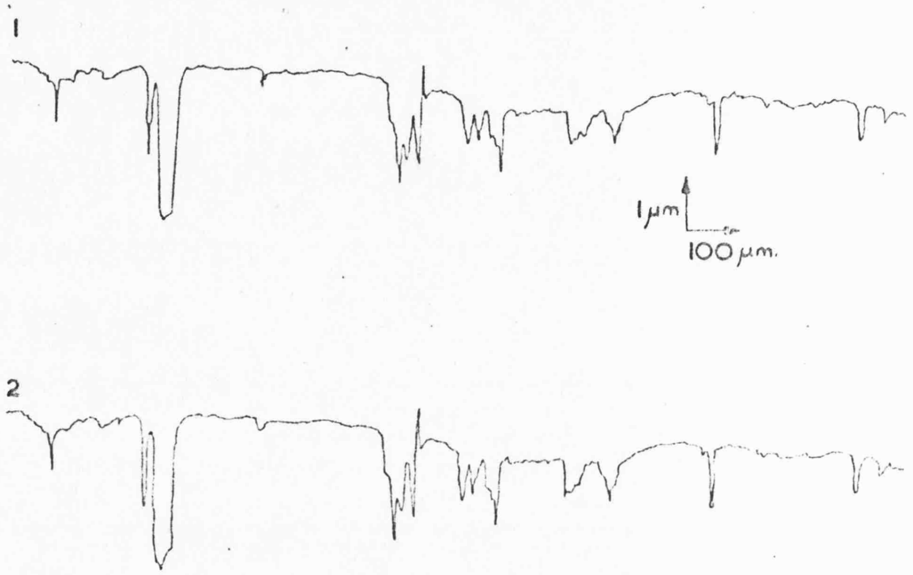
Turning now to the position of the wear track on the specimen. It was found advantageous to have the wear track directly on top of the specimen i.e. in such a way that during sliding the gimbal beam was horizontal. This was checked with a spirit level. Failure to do this produces a wear track not quite parallel to the axis of the specimen.

So far it is clear that many sources of error are possible in an experiment of this nature especially where a high degree of accuracy and repeatability is being sought. Fortunately there was one way in which these problems could be relieved, even assuming that the careful measures that had been taken to ensure accuracy had proved to be insufficient. This method consisted of literally fitting the measurement track to the wear track. This was done by depositing a thin layer of carbon on the lower specimen, making one wear track at a very light load and for a limited extent of the possible traverse. The specimen was then removed from the friction rig and measured in the Talysurf. Using this technique it was possible when viewed under a microscope to distinguish between the marks left on the surface by the stylus and those left by the upper cylinder. The radial stops and the base screws on the vee block were then adjusted until the measured track ran exactly through the centre of the wear track.

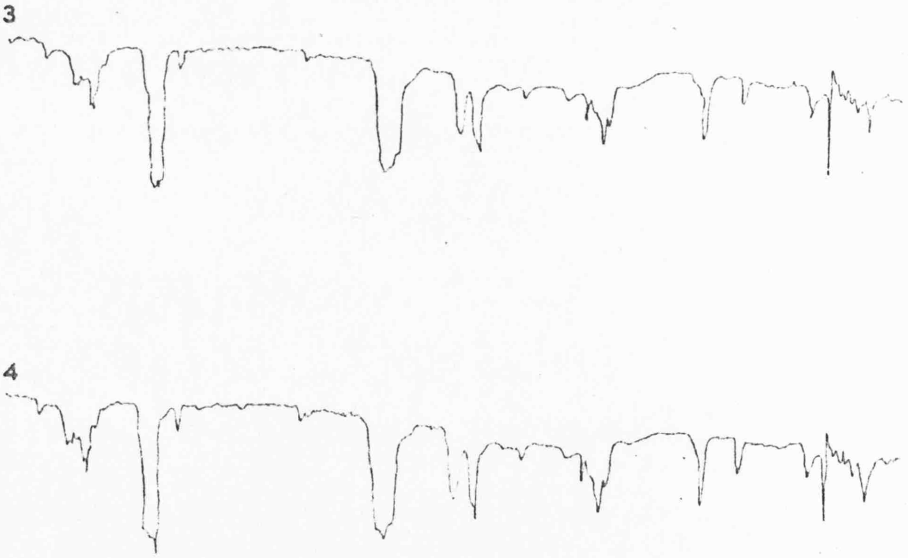
6.2.4 Relocation profilometry - results and other features

For the greatest possible accuracy and repeatability it is essential that neither the vee block nor the Talysurf gearbox,

FIG. 6 –7



RECORDS 1 AND 2 TAKEN ALONG SAME TRACK, RELOCATING SPECIMEN AND READJUSTING MICROMETER BETWEEN TRACKS.



RECORDS 3 AND 4 TAKEN ALONG A TRACK 12μm AWAY FROM TRACK OF RECORDS 1 AND 2. AS BEFORE, SPECIMEN RELOCATED AND MICROMETER READJUSTED BETWEEN TRACKS.

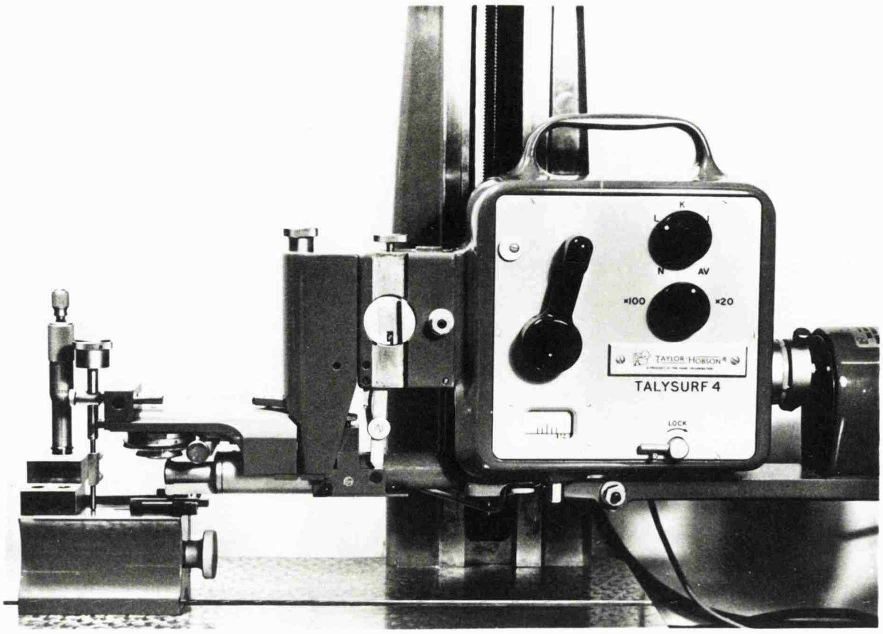


Figure 6-8. Talysurf relocation. This shows the Talysurf and the vee block with the specimen in position.

once set, are moved. Removal and replacement of the specimen in the vee block has to be achieved by slightly raising the pick-up head against its ligament hinges to obtain enough clearance, figure 6-8. Figure 6-7 shows the degree of repeatability achieved which is comparable with that of other workers using different techniques (i.e. Hunt 1968).

A useful facility was introduced which enabled the chart recorders associated with the friction and ride measuring systems to be synchronised. Both could be operated by one switch which enabled accurate determination of the relative phase of the friction and ride graph. This feature enables us to obtain useful results which are described in Section 6.5.

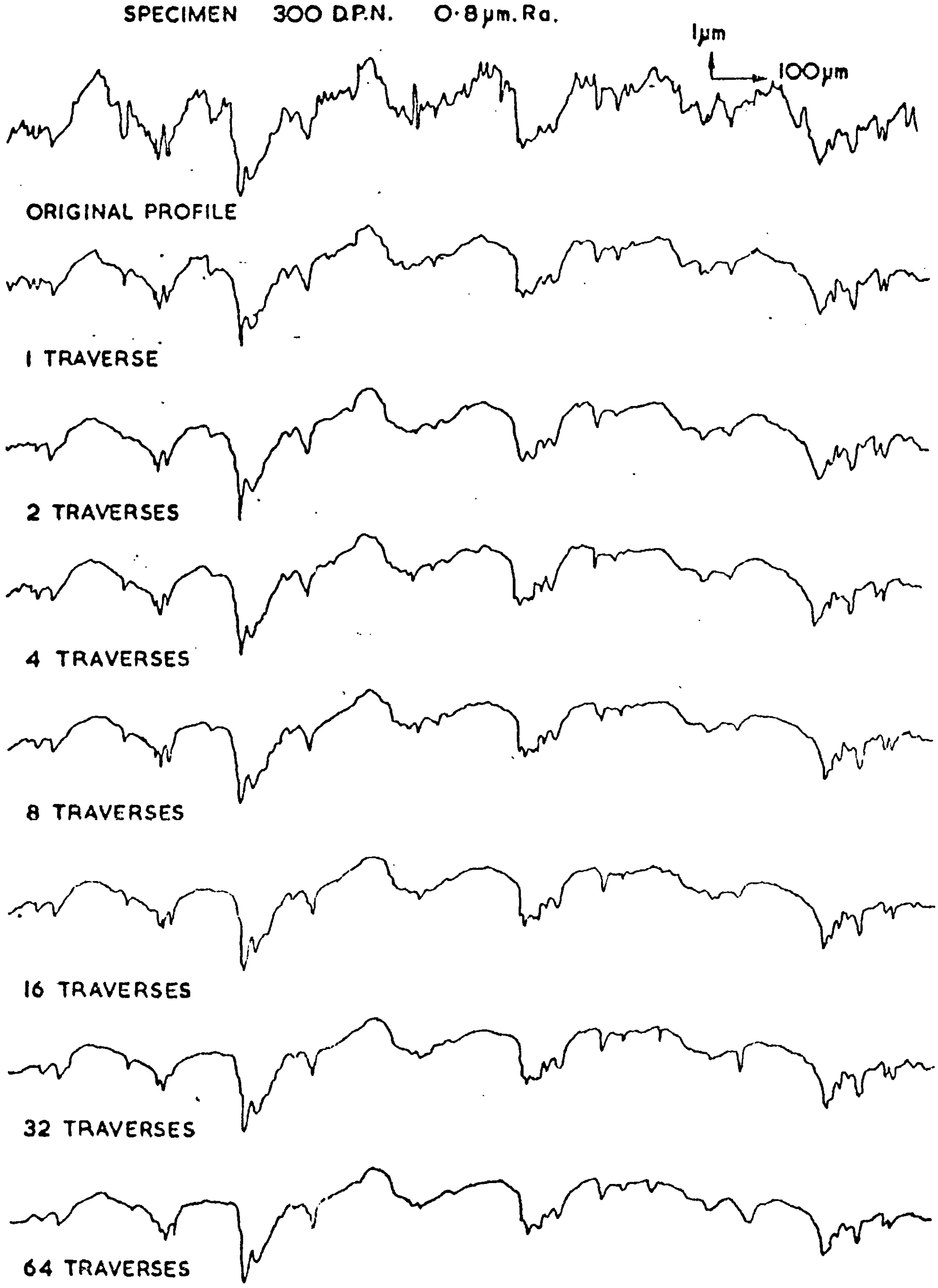
In the consideration of the ride, obviously the errors in the slideway should be taken into account. This is not difficult because over small distances the error curve is small. In fact, because for most practical cases the data logger is coupled to the ride, it would be a simple matter anyway to remove the slideway errors in the computer.

Some work in the development of the machine to make it suitable for the type of experiment described here was carry out by Lunn (1969) as part of a final year undergraduate project. He took steps to overcome vibration in the friction and ride graphs due to the motor and gearbox. He mounted the carriage drive motor and gearbox on a separate framework and used a flexible coupling to the gearbox and the main shaft.

FIG 6-9

THE VARIATION OF PROFILE WITH REPEATED TRAVERSES AT 2·5 Kg.

SPECIMEN 300 D.P.N. 0·8μm. Ra.



ORIGINAL PROFILE

1 TRAVERSE

2 TRAVERSES

4 TRAVERSES

8 TRAVERSES

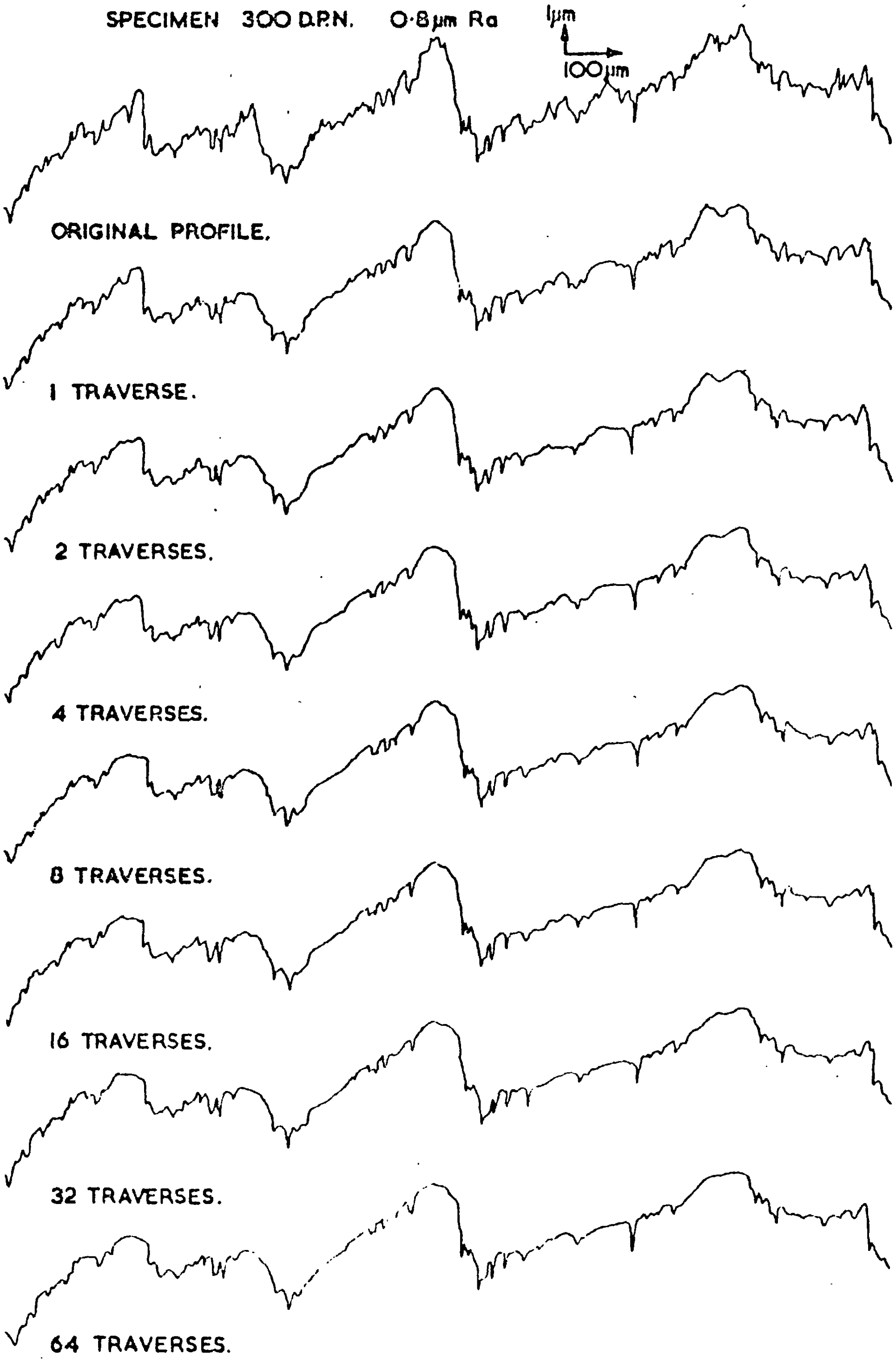
16 TRAVERSES

32 TRAVERSES

64 TRAVERSES

FIG 6-10

THE VARIATION OF PROFILE WITH REPEATED TRAVERSES AT 0·5Kg.



ORIGINAL PROFILE.

1 TRAVERSE.

2 TRAVERSES.

4 TRAVERSES.

8 TRAVERSES.

16 TRAVERSES.

32 TRAVERSES.

64 TRAVERSES.

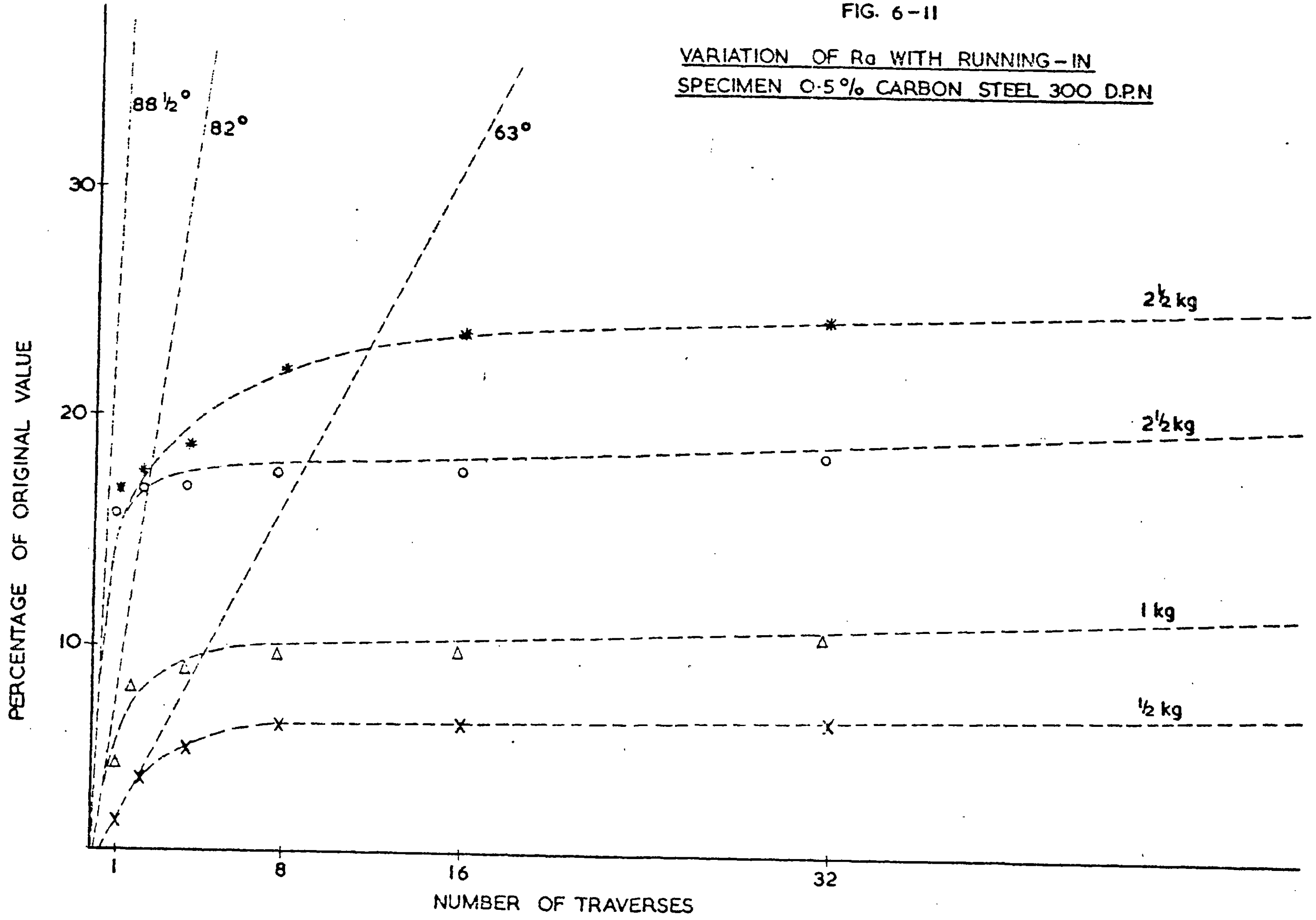


FIG. 6-11

VARIATION OF Ra WITH RUNNING-IN

SPECIMEN 0·5% CARBON STEEL 300 D.P.N

FIG 6—12

THE VARIATION OF PROFILE WITH SINGLE TRAVERSES AT INCREASING LOAD.

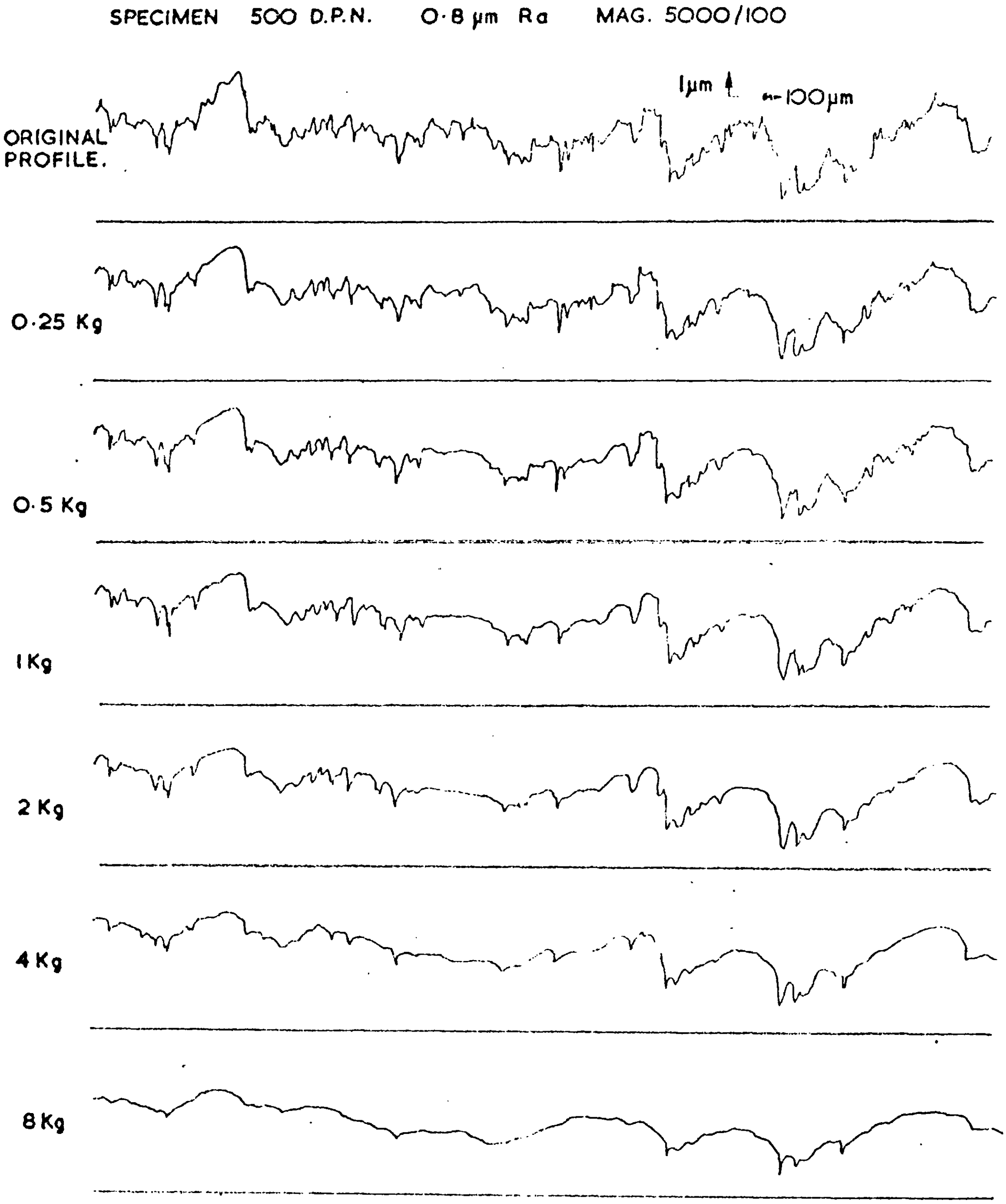
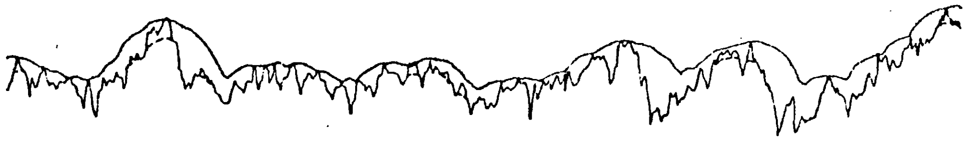
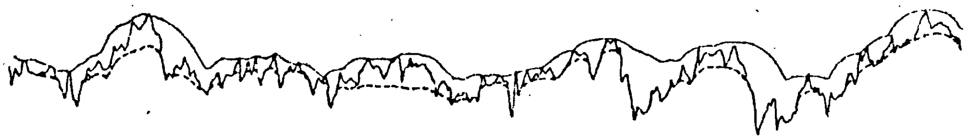


FIG. 6 -13

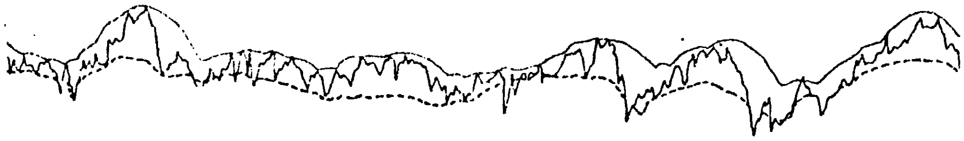
ENVELOPE IS MOVEMENT OF UPPER ON LOWER MEMBER
PROFILE IS OF LOWER MEMBER X 5000 /100



DOTTED LINE IS RESULTANT PROFILE AFTER I TRAVERSE 0 25 kg



DOTTED LINE IS RESULTANT PROFILE AFTER I TRAVERSE 2 kg



DOTTED LINE IS RESULTANT PROFILE AFTER I TRAVERSE 8 kg

6.3 Wear; changes in the surface geometry

6.3.1 Qualitative description

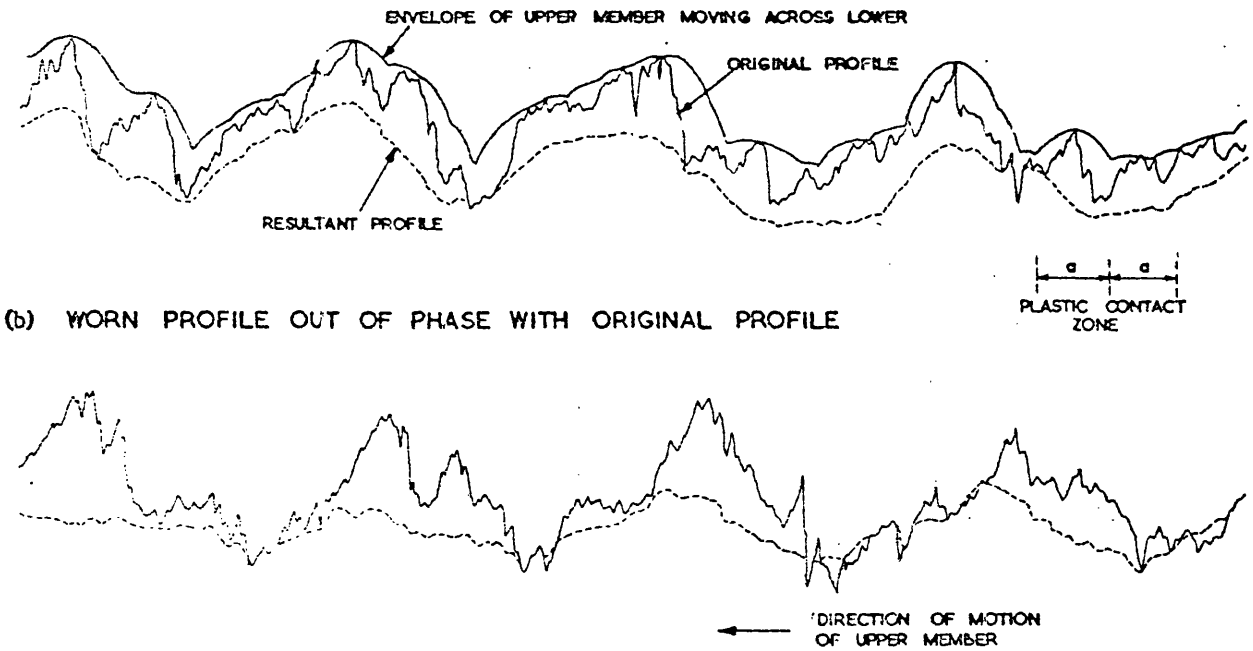
Use of the relocation technique described in Section 6.2.3 enables the geometrical changes of the surface produced by the wear process to be studied in detail. Figure 6-9 shows a typical example of the results obtained using this technique. It illustrates how the surface geometry changes when two cylinders (62 mm diameter) of 0.5% carbon steel of hardness 300 D.P.N. are rubbed together at a fixed load of 2.5 kg under boundary lubrication. The roughness of the cylinders was in this case 0.8 micrometre R_a and the lubricant was a mineral oil. Figure 6-10 shows similar results using a load of 0.5 kg. All these graphs show that it is in the first few traversals of the load that much of the shorter wavelength structure of the surface finish is lost; the long wavelength structure is more likely to be preserved, a point that will be referred to again in Section 6.3.2. As a first simple measure of the changes in topography the change in R_a value is illustrated in figure 6-11 and it will be seen that after the first few traversals further changes are relatively small. Another type of test in which the normal load is increased in successive traversals is shown in figure 6-12. Here again, the same features are displayed.

The accuracy of the relocation technique enables the worn profile to be plotted back onto the original unworn profile both for the case when the profile was worn away with many traversals at fixed load and with progressively increasing load. This is shown in figure 6-13 where the worn part of the profile appears as a dotted line while the full line is that of the original profile (which includes those parts of the worn profile that have been unchanged). It is immediately obvious that it is the smaller peaks situated high up on the profile that have the greatest probability of being removed. Because of the possible presence of adjacent peaks which will carry the load a given peak will not necessarily be smoothed (especially at light loads). The curved line on top of these profiles will be explained later in Section 6.3.2.

Another sort of profile change in the crossed-cylinders type of experiment is shown in figure 6-14(b) which shows just how the metal has behaved at the edge of the wear track. Using the adjustment on the relocation device on the Talysurf enables an investigation to be carried out all over the wear track. In this particular example the trailing edge of the wear track illustrates the "phase change" that has taken place. Here the word "phase" is used spatially and not metallurgically. In the crossed-cylinders machine this phase shift occurs on the sides of the wear track. The reasons for this feature are associated with the way in which one surface moves over the other in the crossed-cylinders arrangement. A given point on one specimen traces out a path on the other at 45^{o} to the tracks which run axially along the lengths of the two cylinders, figure 6-14(c).

FIG. 6–14

METAL FLOW IN WEAR TRACK



(a) WORN PROFILE IN PHASE WITH ORIGINAL PROFILE

(b) WORN PROFILE OUT OF PHASE WITH ORIGINAL PROFILE

MAGNIFICATIONS 5000/100 ROUGHNESS 0·8 $\mu m R_a$ HARDNESS 300 DPN

6.3.2 Wear or running-in as a mechanical filter

Also plotted on figures 6-13 and 6-14(a) is the locus of the lowest point of a smooth upper specimen supposed to have been moved across the original profile without deformation. What stands out in this plot is the remarkable resemblance of the shape of the subsequent worn profile to that of the shape of the locus of the smooth upper cylinder. It appears from this, therefore, that the final shape of the worn or run-in profile is determined to a large extent not only by the original profile but also by how the other body contacts it under no load conditions. This is precisely the situation dealt with in Chapter 5 when the Envelope System of surface measurement was considered.

Two things emerge from this and similar experiments. First, the worn profile and the original envelope become increasingly similar as the running-in proceeds. Second, the wear process is in effect acting as a mechanical filter.

Thus, because the worn profile progressively approaches the envelope in shape as the wear continues to take place and because the average characteristics of the envelope are determined largely by the ratio of the radius of the part to the independence distance, it is not unreasonable to suggest that the correlation distance is a factor of major importance in determining the run-in profile i.e. the correlation distance is of major importance in the wear process.

We consider now the separation of the envelope line from the worn part of the profile shown in figures 6-13 and 6-14. As a first step to determine whether this separation can be explained in terms of the deformation it is useful to decide whether this deformation is elastic or plastic.

Consider first a load of 8 kg shown in figure 6-14. Assuming that the yield stress can be approximated to the hardness which in this case is 300 D.P.N. then the radius of the contact zone under plastic deformation would be given by:

$$a = \sqrt{\frac{W}{\pi H}} \qquad (6\text{-}1)$$

where W is the applied load and H is the hardness; then, from equation (6-1)

$$a \sim 10^{-1} \text{ mm}$$

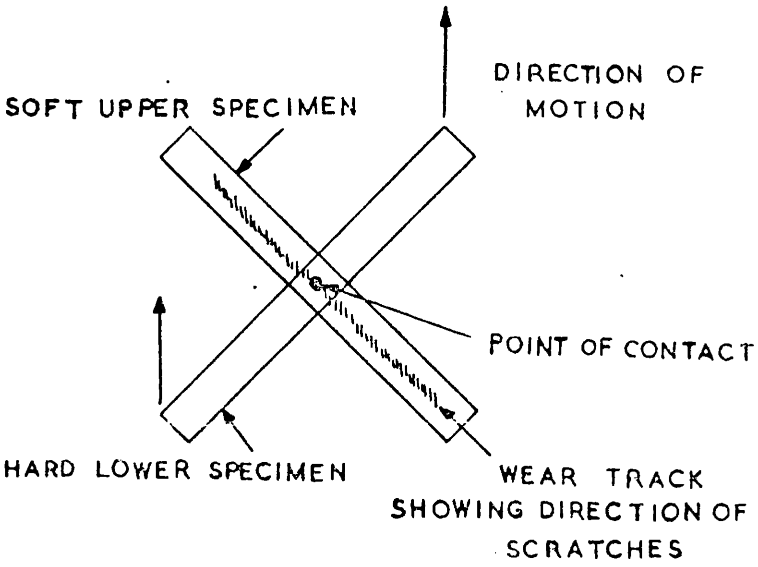
For elastic conditions

$$a = 1.11 \left(\frac{W R}{E}\right)^{1/3} \qquad (6\text{-}2)$$

Where W = 8 kg, E is the elastic modulus = 2 x 10^{12} dynes/cm^2 and R is the radius of the specimen. From this equation $a \sim 10^{-1}$ mm which compares with that for the plastic case hence the condition must be near to critical load, midway between the onset of plastic flow and fully plastic conditions. Thus either equations can be used to give an idea of the separation.

FIG 6-14C

ACTION OF CROSSED-CYLINDERS MACHINE



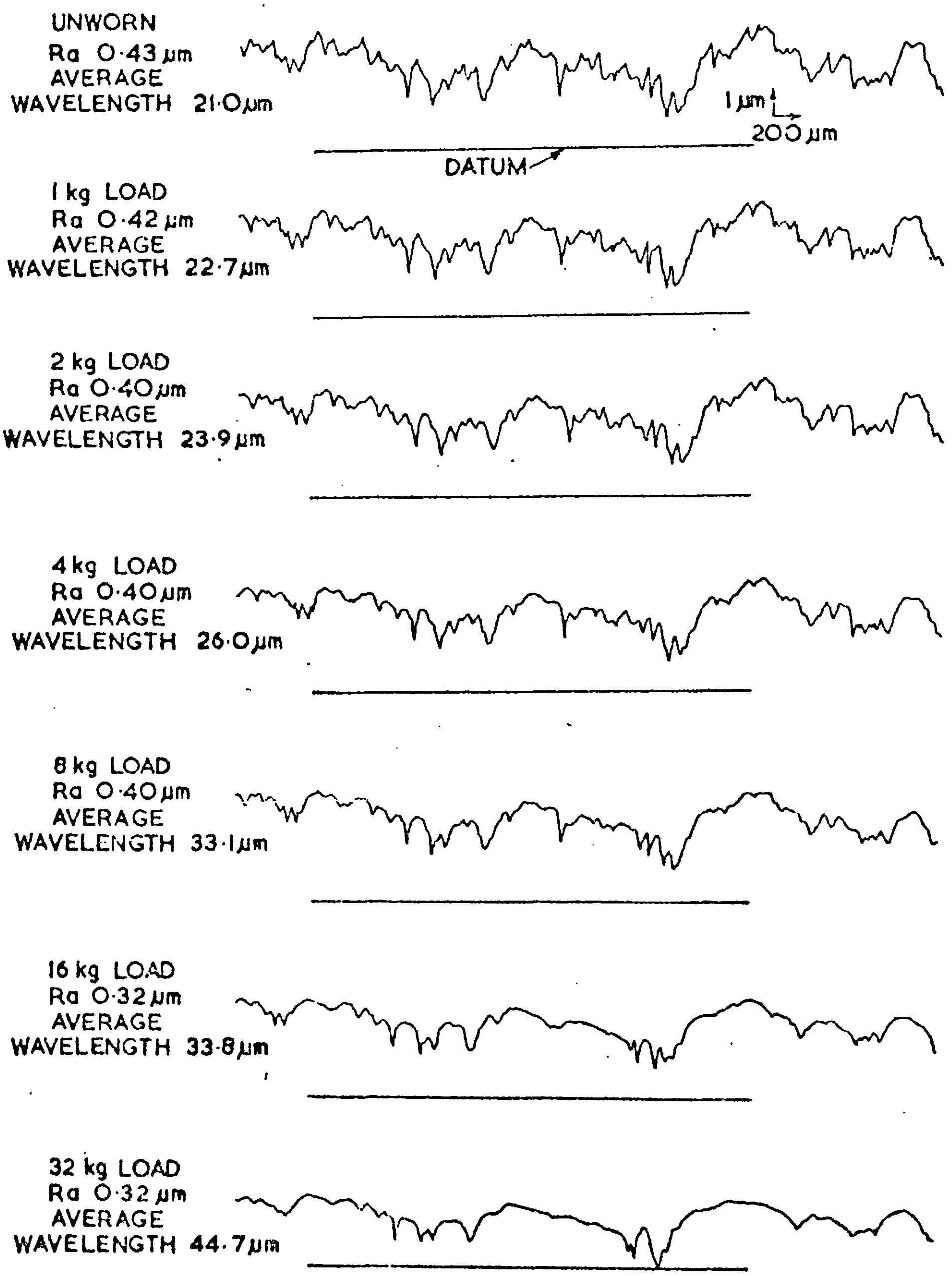
For a spherical contact to be plastically deformed such that the radius of the contact is 10^{-1} mm, using the spherometer formula $h = a^2/2R$, the depression h must be about 1.5 μm. Examination of the graph in figure 6-14 shows an average drop in level of about this magnitude suggesting that the drop in level is indeed associated with deformation.

The fact that the instrumental technique is capable of detecting changes in level of much less than this figure can be seen in figure 6-15 which shows a set of results taken on a ground surface. Special precautions were taken to preserve the d.c. level during the experiment. Apart from the case where the applied load was 32 kg,and well over the plastic limit,the repeatability of the d.c. level is of the order of 0.25 μm. This means that, at least as far as this type of experiment is concerned,the valleys do not apparently move upwards during the running-in process.

This conclusion about lack of movement of the valleys is different from that of Williamson (1968) who showed that in simple loading experiments in which no tangential movement was allowed, i.e. the specimen was completely contained in a "pot", the valleys moved up under extreme loads to the mean plane level at the same time as the peaks moved down. In our experiments, however, tangential movement was allowed and there is evidence that because the metal was not completely constrained as in the Williamson experiment the metal was pushed to the side of the wear track (figure 6-14(c)) and no rise in the level occurred. Figure 6-14(b) shows how, by the use of the micrometer adjustment on the

FIG. 6-15

VARIATION OF Ra AND AVERAGE WAVELENGTH WITH RUNNING-IN



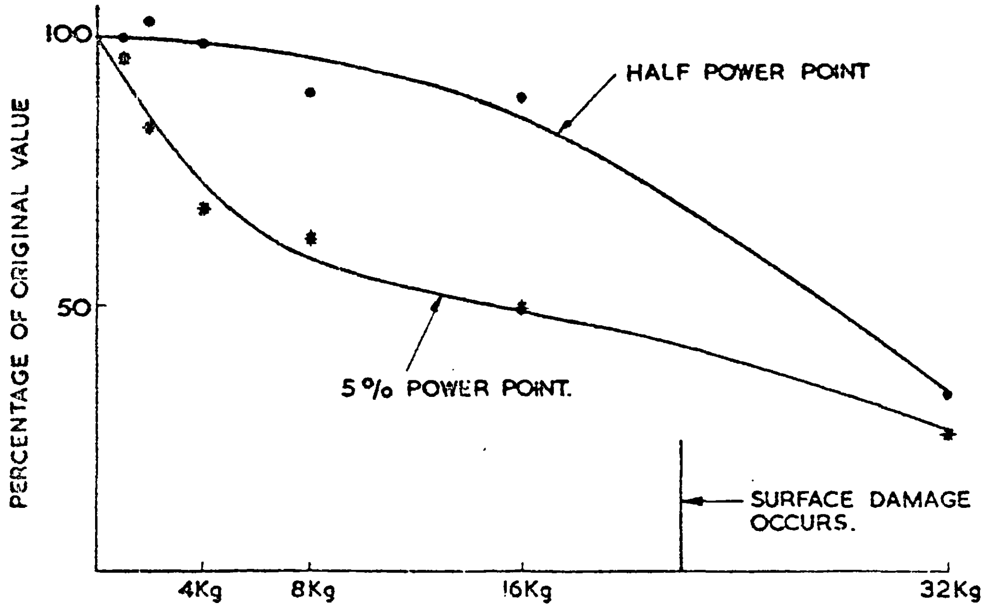
relocation device, the metal flow in the wear track can be investigated. It should be possible, using this technique and taking many parallel traces across the wear track, to find out exactly where the displaced material goes.

6.3.3 Parameters of the profile and their modification by rubbing

It has already been stated that surface geometry parameters likely to be of importance in the wear process are the curvature at the peaks and the slope of the flanks because according to modern theory these factors of the geometry, combined with the mechanical properties, determine whether the peaks will be elastically or plastically deformed. Obviously the material displaced by plastic deformation will cause the geometrical changes during the wear process. The relocation apparatus, together with the data logger and the crossed-cylinders machine, provide the opportunity to test the idea that these features of the surface topography are those most directly affected by rubbing. It might be expected that the effect of running-in will be that the curvature or slope of the asperities will be rapidly adjusted by plastic flow so that after a few traversals the deformation becomes primarily elastic, figures 6-9 and 6-10. Measurements of these sophisticated parameters can be carried out during the wear run by means of digital techniques - measurements which would be very difficult by any other means.

FIG 6-16

CHANGES IN POWER SPECTRUM PARAMETERS DURING RUNNING IN.



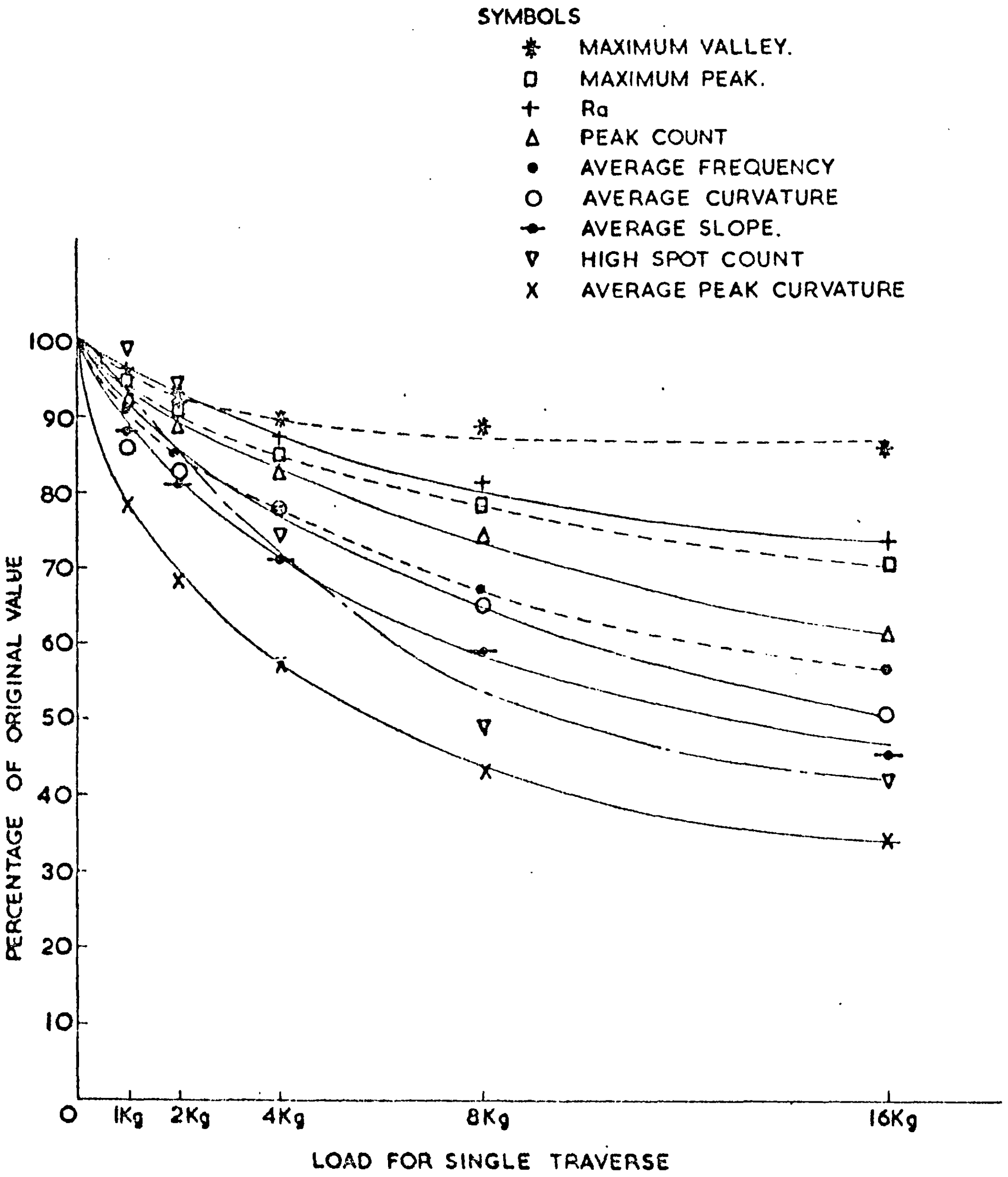
Soft specimens were deliberately selected (300 D.P.N.) so that the geometrical changes could be expected to show up quickly. As before the upper member was a hard, polished cylinder. The original R_a was typically 0.5 μm. In addition to the profile both the friction and the "ride" of the upper member on the lower were recorded in digital form on punched paper tape. Itshould be noted that failing the availability of recording apparatus or two data loggers these could not be recorded simultaneously.

As well as the two parameters mentioned above being digitally measured, many other types of parameter could be evaluated from this basic data; for these some special programs had to be written (Appendix 2). Because all the parameters are measured from the same sets of data then a truly meaningful comparison can be carried out. The wear process was accomplished by performing runs at progressively heavier and heavier loads. Between runs the profile was digitally recorded.

Figure 6-16 shows how the power spectrum of the surface changed during the wear process. It shows clearly that the components of the geometry carrying most of the "energy" (taken as those having wavelengths longer than the half power point) were hardly affected until the load applied was severe enough to cause considerable plastic deformation; however, the small wavelength components, i.e. those less than the five percent power point, were considerably attenuated with even the smaller loads. This provides some direct experimental support for the qualitative argument advanced above that one major effect of rubbing is that the broad

FIG 6-17

VARIATION OF PARAMETERS WITH RUNNING IN.



scale structure tends to be preserved. Thus we would expect that the run-in surface geometry might be described quite satisfactorily by the zero-order Markov chain model of the surface; the higher frequencies not having anything like the same importance.

A comparison of some of the other parameters measured is shown in figure 6-17. In this all the parameter values for the original profile have been taken as unity to display more clearly the changes that occur during running-in. A result which can be observed immediately is that the average curvature at the peaks changes more rapidly with rubbing than does any of the other parameters. The average slope also is very sensitive to wear, whereas the R_a value commonly used in surface finish is relatively stable, dropping to 73% of its original value as opposed to less than 35% for the average peak curvature. These results show that it is, indeed, peak curvature and slope that change most in the wear process; this is the rcsult forecast at the start of this section based upon the ideas of surface contact discussed in Chapter 4.

Figure 6-18 illustrates the close similarity of results which are obtained by measuring either average or root mean square values. Because of this, and for simplicity, average values have been used to display the changes in figure 6-17. These results are only taken up to a load of 16 kg because the surface became torn for loadings much higher than this.

To emphasise the role of the wear process as equivalent to that of a high-cut filter consider a comparison of figures 6-17,

FIG. 6-18

VARIATION OF PARAMETERS WITH RUNNING IN - AVERAGE & R.M.S. VALUES.

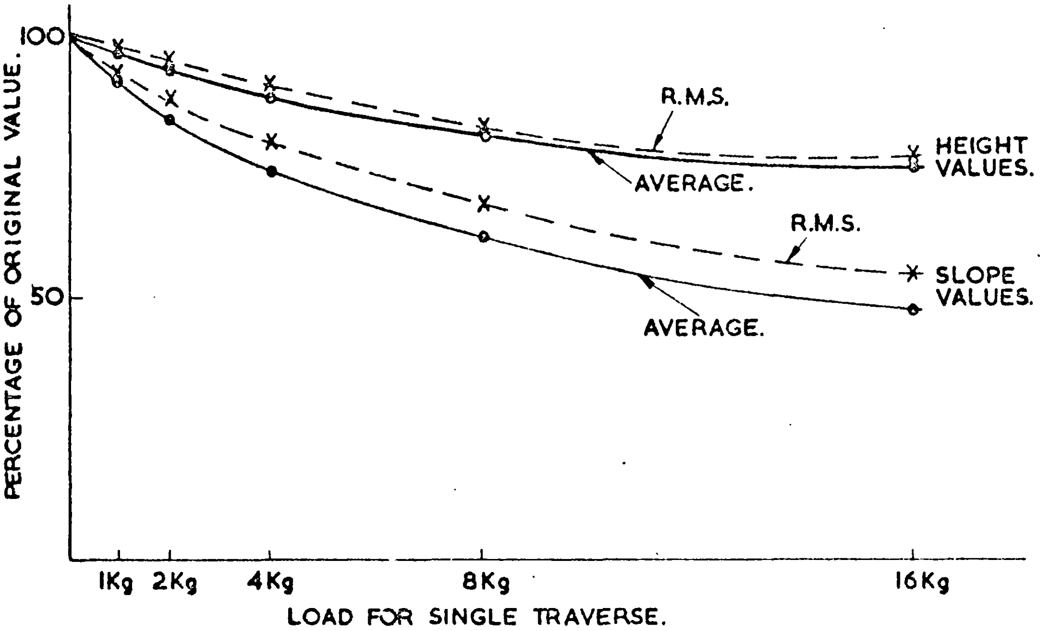
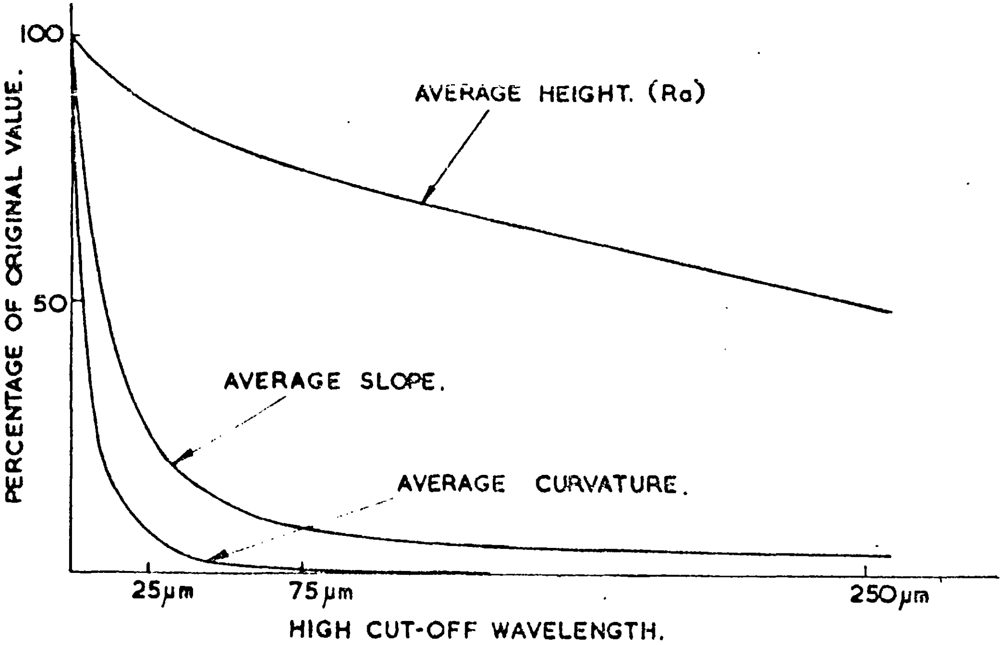


FIG. 6-19

VARIATION OF PARAMETERS WITH DIGITAL HIGH CUT FILTERING.
(STANDARD 2CR FILTER.)

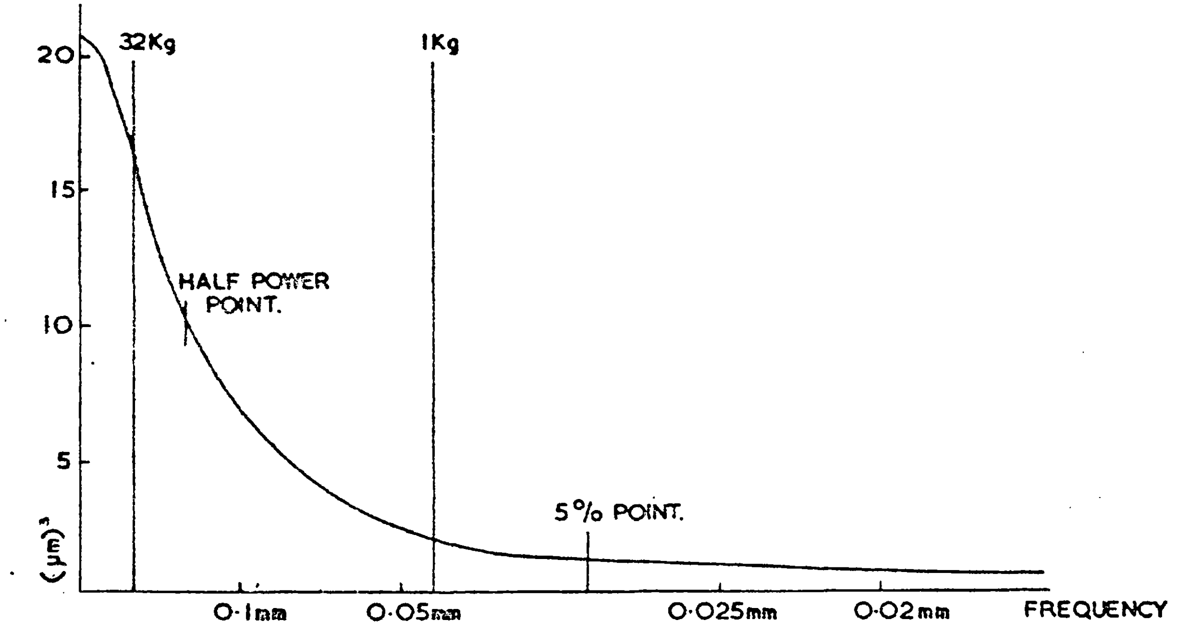


6-18 and 6-19. Figure 6-19 shows how three parameters, the height, slope, and curvature, of the profile change when the waveform is subjected to a high-cut digital filter of 12 dB per octave. Because, by their very nature, curvature and (to a lesser extent) slope are dominated by the short wavelengths they tend to become dramatically attenuated as the filter cut-off is brought into the wavelengths containing most of the energy. On the other hand, the average height (R_a) does not change rapidly with changes in high frequency cut. In effect, as a broad guide drawn from the comparisons of figures 6-17 and 6-19 it might be suggested that the running-in process for this specimen hardness as typified up to loads of 16 kg is equivalent, as far as the R_a value is concerned, to a high-cut filter cutting at wavelengths of 75 μm whereas the equivalent cut is at about 16 μm for slopes and 5 μm for curvatures. Care must, however, be exercised in extending the comparison between the effects of filtering and wear. Filtering affects the whole of the profile; as the results obtained in this chapter show, but the influence of wear, particularly at light loads, is upon the upper part of the waveform.

Using equation (6-1) some idea of the extent of the filtering effect of the wear process can be obtained. This shows that assuming that the crossed-cylinders contact can be approximated by the contact between smooth surfaces, the width of the plastic zone is approximately 0.06 $\sqrt{W}$ where H has been taken as 300 D.P.N. and W is in kilogrammes. Equation (6-1) also shows that the plastic zone radius is inversely proportional to the square root of

FIG. 6-20

POWER SPECTRUM AND ITS RELATION TO PLASTIC ZONE SIZE.

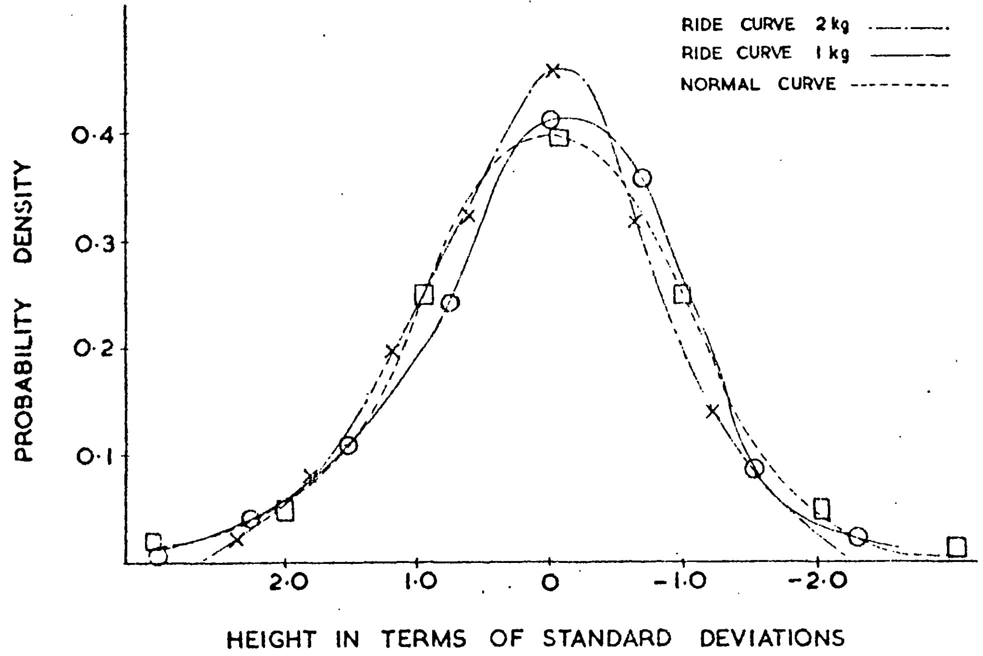


hardness and hence the effective filter cut-off is likely to be related inversely to hardness. Figure 6-20 shows where these zone widths lie with respect to the power spectrum. It illustrates why the five percent power point is more strongly affected by small loads than is the half power point. However, it should be emphasised that the zone widths drawn on this graph cannot be compared directly to that of the cut-off of a digital or electrical filter. This would imply that the zone averages the profile in a way similar to a rectangular weighting function; this cannot be so because the wear process is inherently asymmetrical. But the example does serve to illustrate the shape of the curves in figure 6-16.

Some other points concerning the way in which the profile changes during wear should be mentioned. The feature of asymmetry, just described, obviously produces an asymmetric change in the profile geometry. This can be demonstrated by measuring the skew of the profile distribution or some derivative of it, for instance, the maximum peak minus the maximum valley divided by the R_a value. Changes of two to one over the same range as in figure 6-17 would result. Unfortunately, skew as well as the other high order moments of the ordinate height distribution are not very stable measures, being strongly influenced by rare events in the sample. One further point is that all measures of peak or ordinate height are not very much changed by the wear process. Any measure of the valley height is affected even less. This does not necessarily mean that these are poor parameters; it depends whether the parameter

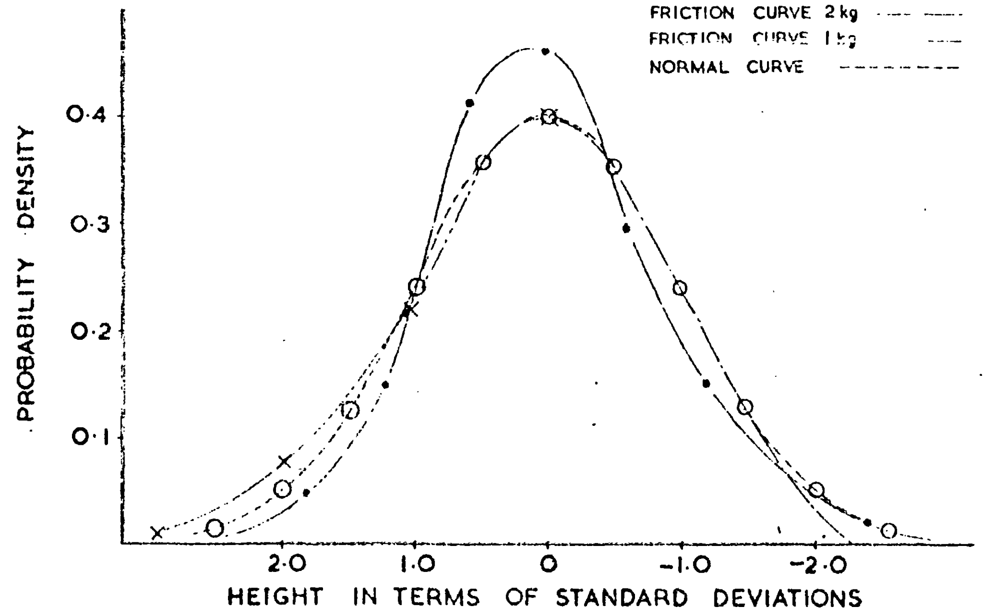
FIG. 6 -21a

DISTRIBUTION OF RIDE CURVE



(b)

DISTRIBUTION OF FRICTION CURVE



is being used as an estimate of the surface finish which will result after the running-in stage has been completed. For example, the original R_a value, which does not change much during the wear process, is often used as an estimate of the run-in value.

6.4 Movement of rubbing bodies under load; the ride

As mentioned in Section 6.2 the part of the crossed-cylinders machine which measures the ride consists of a pick-up mounted in such a way as to pick up the vertical movement of the upper member in its travels across the lower one. Using this apparatus it has been possible to measure the ride of the upper specimen for various loadings and hence come to some conclusions about the elastic and plastic properties of the contact and in particular elastic recovery. Ride is, however, quite difficult to measure except for really rough specimens. This is because of the inherent errors in the carriage slideway and the tendency of the apparatus to deform under load. Consequently considerable digital filtering may be required to extract any useful information from the waveform. For a rough specimen in which the carriage errors are small compared with the ride the situation is much simpler.

The nature of the ride waveform can be seen clearly in figure 6-21(a) which shows how nearly the curve approximates to a Gaussian distribution. Any differences can no doubt be attributed to the fact that the length of sample is statistically small; because the bandwidth of the ride waveform is smaller than that of the corresponding profile one requires a longer duration to get the same degree of reliability as that of the equivalent profile. This statistical difference is also illustrated by the differences that occur between the 1 kg and 2 kg curves which are both shown. An application of the ride will be discussed in Section 6.5.2.

6.5 Friction

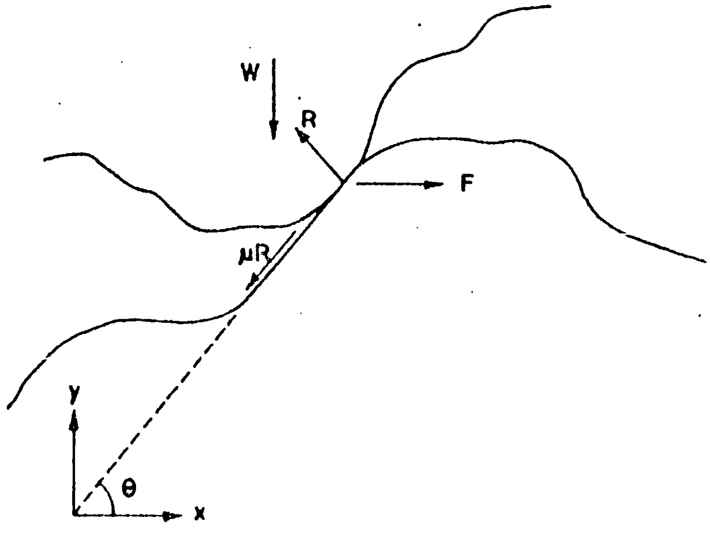
6.5.1 Introduction

The earliest systematic work was done by Coulomb (1785) who was the first to give an analytic approach. He suggested that friction between surfaces was principally due to the work required to lift loaded asperities over each other. One of the reasons why this approach has met with little acceptance in modern times is because the mechanism is not dissipative; it could not, on its own, explain a steady frictional force.

Recent work has pointed to two major mechanisms being mainly responsible for the frictional force; the so-called adhesion term which is related to intermolecular forces within the surface layers and a ploughing term which is necessary to explain the, sometimes irreversible,deformations caused in the bulk of the material by the ploughing effect of the asperities. In what follows some experiments will be described which try to reveal some aspects of the friction and how it is influenced by surface finish. The experiments were carried out at slow speeds (10 μm/sec and 50 μm/sec) so as to avoid heating effects. A small amount of oil was present so that boundary lubrication conditions prevailed.

FIG. 6-22

COMPONENT OF COULOMB FRICTION



Only the fluctuating components of the friction were investigated because these were considered to be more related to the surface finish than the mean value which although usually much larger, is more related to the adhesive term.

6.5.2 Coulomb friction

To see what sort of relationship to expect between the frictional force and the surface geometry consider figure 6-22 which shows the situation where two asperities have interlocked. Let μ be the coefficient of friction due to adhesion and $\overline{F}$ the force required to pull the one asperity in the direction shown. Also let the upper surface have a load W on it causing a reaction $\overline{R}$ at the point of contact. Assume further that the average angle at the point of contact is θ. Resolving vertically

$$W + \mu \overline{R} \sin \theta = \overline{R} \cos \theta \qquad (6\text{-}3)$$

horizontally

$$\overline{F} = \mu \overline{R} \cos \theta + \overline{R} \sin \theta \qquad (6\text{-}4)$$

from which

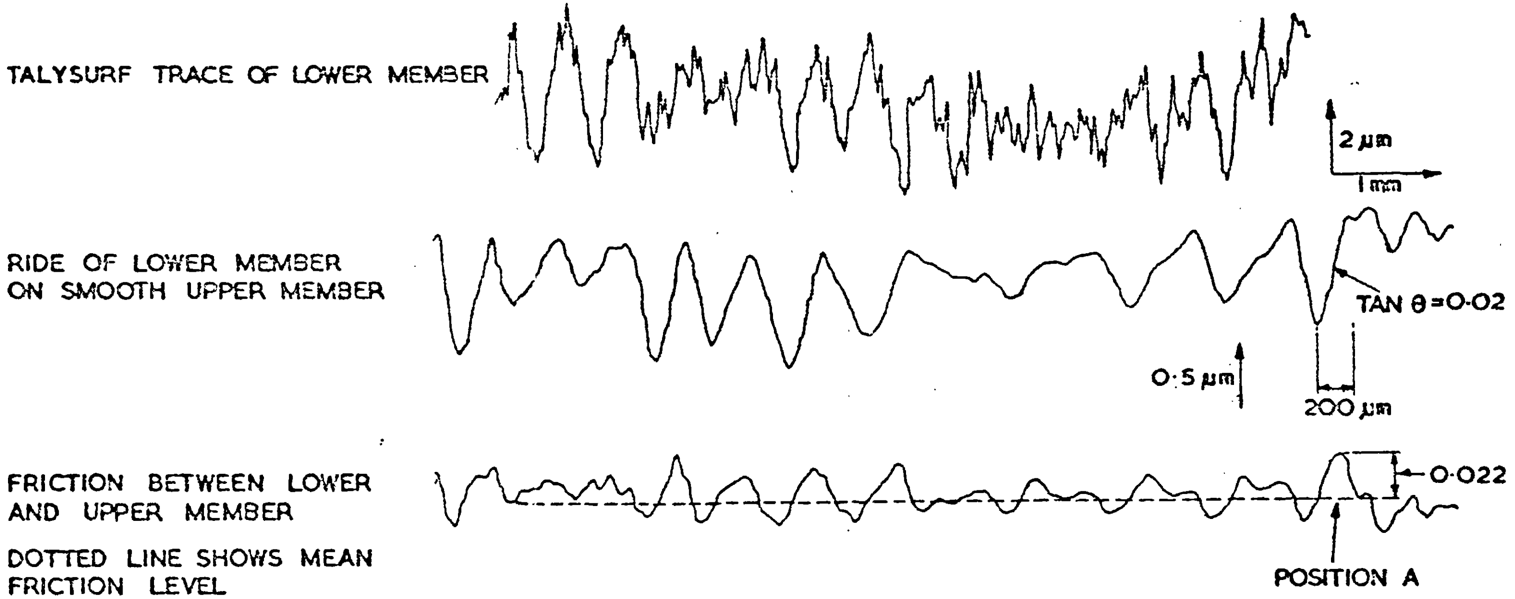
$$W = \overline{R} \cos \theta - \mu \overline{R} \sin \theta \qquad (6\text{-}5)$$

The coefficient of friction actually measured is $\overline{F}/W = \mu_{eff}$. Hence μ_{eff} is given by the division of equation (6-4) by equation (6-5) from which

FIG. 6-23

RELATIONSHIP BETWEEN CCULOMB FRICTION AND SURFACE GEOMETRY

UPPER SPECIMEN SMOOTH
LOWER SPECIMEN O·8 μm Ra
O·5 % CARBON STEEL
BOUNDARY LUBRICATION



LOAD ON SPECIMEN 2 kg - LOAD ON BEAM 5OO grms
FRICTION CALIBRATION 2OO grms GIVES 3O mm ON
GRAPH AT 5OOO x

$$\mu_{eff} = \frac{\mu\,\bar{R}\cos\theta + \bar{R}\sin\theta}{\bar{R}\cos\theta - \mu\,\bar{R}\sin\theta} \qquad (6\text{-}6)$$

$$= \frac{\mu + \tan\theta}{1 - \mu\tan\theta} \qquad (6\text{-}7)$$

For the case of real surfaces $\tan\theta < 0.1$ and $\mu \sim 0.1$ so that $\mu\tan\theta$ can be neglected compared with unity, hence

$$\mu_{eff} = \mu + \tan\theta$$

or $$\mu_{eff} = \mu + y' \qquad (6\text{-}8)$$

In equation (6-8) $\mu = \underline{S}/H$ where $\underline{S}$ is the average shear strength of the area of contact of the two surfaces and H is the microhardness of the surface layer. Equation (6-8) says, in effect, that measurement of the average level of friction gives a measure of the adhesion and if this is assumed to be constant along the surface then measurement of the fluctuating friction component gives an estimate of the differential of the surface geometry, or at least that part of the geometry which is made apparent in the ride of the one component on the other.

That this is so can be seen in figure 6-23 which shows the variation of frictional force measured at precisely the same time as the ride. In producing these results, synchronisation of the friction record with that of the ride is essential. It can be seen that where the fluctuating component of friction cuts the mean friction level that the ride is at a stationary point. As an example of the quantitative accuracy of this differentiation consider position A marked on the figure. It has a peak value of 0.70 cms. The calibration of the friction graph was such that 200 grams on the end of the beam gave 30 mm on the graph at a magnification four times smaller. This means that the frictional force at the end of the beam corresponding to position A was 11 grams which gives a coefficient of friction of 0.022 in this case, because the normal load was 500 grams. Above this peak on the friction graph is a steep slope of the ride graph. This has a maximum slope, corresponding to the stationary point on the friction graph, of value 0.02.

From this it can be seen that within the accuracy of the experiment the variation of the friction graph does give the differential of the ride graph. Above the ride graph is shown a Talysurf profile trace taken over the same region on the specimen before the friction run had been taken. Although it is difficult to compare this directly with the ride and friction graphs because of the slightly different horizontal magnification, the figure does show a reasonable agreement between the ride and profile.

Whether or not the differentiation can be seen depends on the constancy of μ in equation (6-8). In these experiments,because boundary lubrication exists,the value of μ is smaller than would be expected with dry conditions thus making the ratio of the alternating-to-mean value of friction rather larger than would otherwise be expected.

The results shown here point to the fact that it is the slope of the surface geometry rather than the profile itself which contributes to the alternating components of the frictional forces.

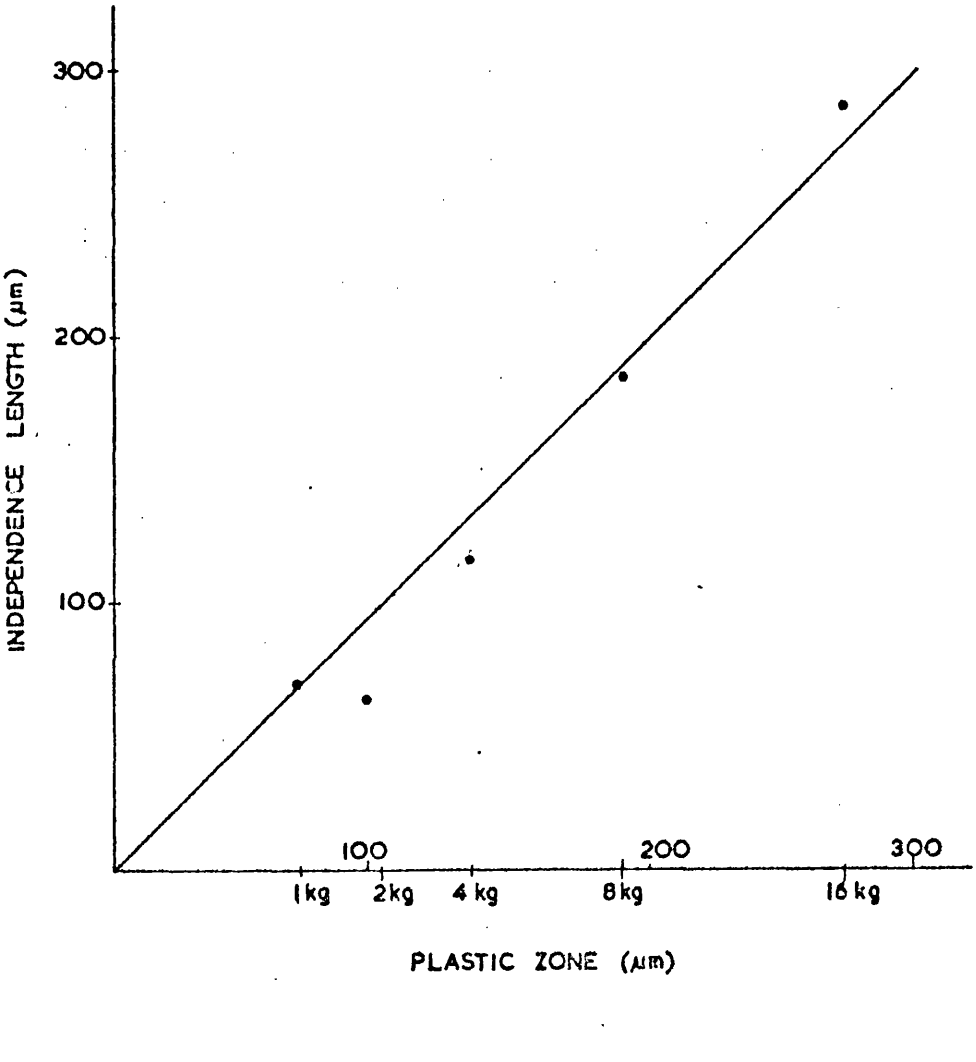
6.5.3 Nature of the frictional fluctuation

To obtain information about the frictional variations the use of the data logging techniques described in Chapter 3 is essential. Well over three thousand measurements of the friction and the ride waveform were taken just for any one track. From these measurements much information can be gleaned. For instance, it was found that, as in the case of the ride (figure 6-21(a)) the friction measurements (figure 6-21(b)) also have a distribution which is very close to Gaussian. This is a point noticed by Nagasu (1951) for static friction and Rabinowicz et al (1954). Again, as in the graph of the ride variation, two examples are given, one at 1 kg and one at 2 kg to give some idea of the variation that can exist between different tracks.

One point concerning the processing of friction information is that the large variations that can sometimes be introduced into the ride waveform due to errors in the apparatus are less likely to be of such importance in friction; in many cases the friction waveform does not have to be processed by a digital low-cut filter before evaluation. Rabinowicz (1957) has proposed measuring the variations in the transverse friction force rather than the ordinary variations in the direction of traverse in an attempt to remove the large average friction value. It has rarely been possible to leave out

FIG. 6 -24

INDEPENDENCE LENGTH OF FRICTION PLOTTED AGAINST PLASTIC ZONE WIDTH



the filtering stage in the case of the ride waveform.

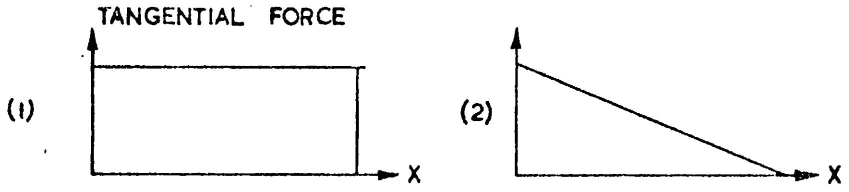
Using the data obtained in digital form from the friction measuring system it has been possible to measure autocorrelation functions of the friction for different loadings. From these autocorrelation functions, which were generally exponential in form, the independence length of the friction has been obtained in much the same way that the independence distance of the profile was determined.

Values of this friction independence distance have been plotted in figure 6-24 for loads on the specimen from 1 kg to 16 kg. This distance is plotted as a function of the width of the plastic zone to be expected on the lower specimen using equation (6-1). From the graph it appears that the distance of independence of each friction waveform under these conditions correlates well with the width of the plastic zone for the load used in each experiment. Remember that in this experiment the upper specimen was smooth and hard compared with the lower specimen.

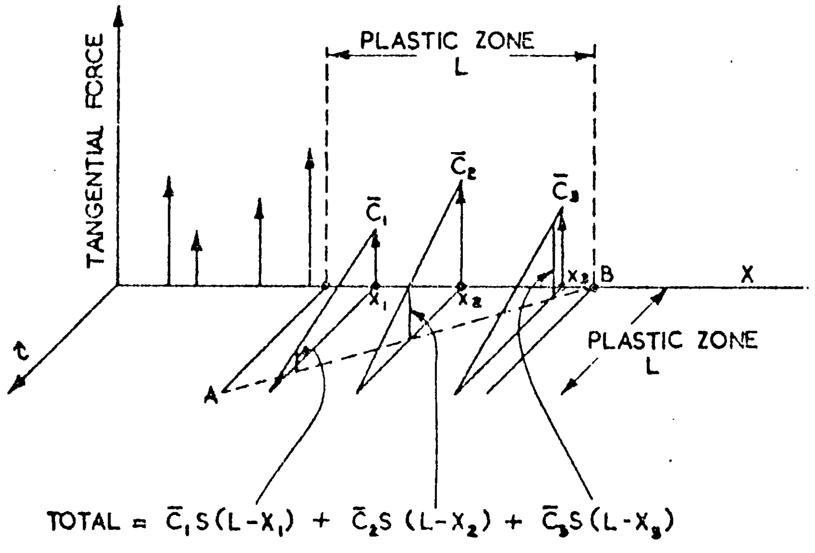
The plastic zone region determines the area within which all the forces making up the frictional force originate. This force will be made up of contributions from all the individual contacts actually within the zone at any one time. Therefore, consider the simplified situation shown in figure 6-25(a). This shows two possible ways in which the tangential force might vary for a single contact from the time that it enters the contact region until the time that it emerges, crushed to some extent, at the other end.

FIG. 6-25

(a) TANGENTIAL FORCE DISPLACEMENT CURVE FOR SINGLE CONTACT IN PLASTIC ZONE REGION



(b) TOTAL TANGENTIAL FORCE DUE TO NUMEROUS CONTACTS IN PLASTIC ZONE REGION



If it is assumed that the effect of the physical size of the contact is merely to scale up the curve in height then the situation, in general, for the simplified two-dimensional case would look like figure 6-25(b) which shows, for clarity, only three contacts $\overline{C}_1$ $\overline{C}_2$ and $\overline{C}_3$ in the contact zone. In this figure τ is a dummy space variable. Again, for clarity, the second shape of the tangential force-displacement curve is shown in the figure 6-25(b) because it enables the actual force being generated by each contact in the zone to be seen more clearly than it would for the first shape. In the method of presentation shown the actual total force is found by adding up the contributions from each contact in the zone along the line AB.

To see why the friction independence distance and the plastic zone should be related consider the situation shown in simplified form in figure 6-25(b). Along the x axis are a number of impulses representing the surface asperities on the lower member; these impulses only refer to those large asperities which are potential contacts. Notice that these asperities likely to be contacts are probably dictated by the independent events of the surface. For a random surface these impulses will be randomly distributed throughout space. If the tangential force-displacement curve for each individual contact within the contact zone is s(x) then the total frictional force that would exist if all these peaks were in contact would be $\overline{M}(x)$, but in practice because of the finite extent of the plastic zone region only a finite number of contacts will actually be contributing to the total friction force at any one position x of the zone on the surface, i.e. the frictional force actually measured at (x), M(x), will be, in effect, a sample from the total population $\overline{M}(x)$; simply, for this exercise, M(x) can be taken to be the convolution of C(x) with the plastic zone width (or more strictly with the tangential force-displacement curve which only exists over the plastic zone width); where C(x) is the function describing the contact distribution in space.

$$M(x) = C(x) * (U\,(x + L/2) - U\,(x - L/2)) \qquad (6\text{-}9)$$

where U(x) is the Heaviside function.

If $\underline{F}(\omega)$ is the Fourier transform of M(x),

$\overline{\underline{F}}(\omega)$ of C(x) and $G(\omega)$ of $(U\,(x + L/2) - U\,(x - L/2))$

then

$$\underline{F}(\omega) = \overline{\underline{F}}(\omega) \times G(\omega)$$

For this simple case

$$G(\omega) = \frac{2 \sin \omega\, L/2}{\omega} \qquad (6\text{-}10)$$

Hence the power spectrum of $\underline{F}(\omega)$, $\underline{P}(\omega)$ is given by

$$\underline{P}(\omega) = \overline{\underline{P}}(\omega) \times \left|G(\omega)\right|^2 = \frac{4\overline{\underline{P}}\,(\omega)\,\sin^2 \omega L/2}{\omega^2} \qquad (6\text{-}11)$$

Where $\overline{\underline{P}}(\omega)$ is the power spectrum of $\overline{\underline{F}}(\omega)$.

Now if, as is drawn in figure 6-25(b), the contacts can be considered as small compared to L in width, then each contact can be considered to be an impulse. Hence $\underline{\overline{P}}(\omega) \rightarrow 1$ providing they are randomly spaced. (Papoulis 1967). Consequently taking the inverse transform of equation (6-11) yields the autocorrelation function C(β)

$$C(\beta) \quad = \quad (1 - \frac{|\beta|}{L}) \qquad\qquad (6\text{-}12)$$

which is triangular and zero at L.

The equation (6-12) assumes that the tangential force-displacement curve is rectangular as shown in figure 6-25(a). This may well not be the case. It could be many different shapes (Rabinowicz 1956, Green 1955) but it will still not affect the basic argument that C(β) will decline towards zero as $\beta \rightarrow L$. This is because, in effect, the graph of the tangential force-displacement curve of the contact acts as the impulse response of the friction behaviour. The fact that it is L in length no matter what it has for a shape will make $C(\beta) \sim 0$ at $\beta = L$. Obviously for the more realistic case where the contact size is not small compared with L then the independence distance of the friction will be correspondingly larger, by the contact size amount. It should be possible from the transform of the autocorrelation function to get some information concerning the true nature of the tangential force-displacement curve. The result shown in figure 6-24, together with the reasoning given above, point to the conclusion that under the conditions of this experiment the independence length of the friction waveform should correlate closely with the plastic zone size. Indeed, under the conditions of this experiment it would appear that the Fourier transform of the tangential force-displacement curve represents an approximate transfer function of the friction process, where input is the profile contacts and the output is the frictional force.

6.6 Discussion and Conclusions

In this chapter the relationship between the topography of surfaces and some of their tribological characteristics has been examined. Naturally, this examination has been exploratory and has been based upon a few simple, but highly instrumented, experiments. The outstanding conclusion which can be drawn from this work is that there is a wealth of information which could be obtained from a more extended series of experiments.

Most of the merits of the work described here, and some of its disadvantages, arise from the use of the crossed-cylinders machine. By its use it has been possible to correlate a given feature of the surface topography. Two important features of the crossed-cylinders arrangement are relevant in this discussion. First, the arrangement provides a single, well-defined, region of contact but a region whose characteristics are affected by the surface topography. Second, in sliding, this region of contact moves continuously to a fresh region of <u>both</u> the specimens; thus any changes in topography and their relation to, say, the measured friction, are not masked by the fact that some points on the surfaces are rubbed many times before changes in topography can be examined. The major disadvantage of the present machine arises from the inaccuracies of the slideway which have contributed to difficulties in making meaningful and significant measurements of the ride, especially in measurement of its autocorrelation function. However, these difficulties arise from the construction of the apparatus, (the only one available when the work commenced) rather than from the principles involved in its design.

Other instrumental features that have been of major significance in this work as distinct from others in the field have been not only the use of relocation techniques to monitor wear, friction and ride but also the use of data logging techniques which have led to a far more precise measure of the parameters during running-in than that which has previously been possible. This has not simply been due to the data logging itself but it is also due to the use of numerical analysis techniques described in Chapter 3. It is the use of these three features simultaneously (a) crossed-cylinders, (b) relocation techniques and (c) digital analysis of results, that have made the overall method so powerful. These have helped to define more precisely the conditions of operation of the experiments and hence give more weight to the results obtained from those experiments.

The difficulties involved with working with undefined conditions applies,therefore,with even greater force to the limited number of earlier published accounts of work in this field,the most important of which will be mentioned below. Experiments have been carried out by Ostvik and Christensen (1968) on the changes in the surface geometry that occur during the running-in process of disks under E.H.L. conditions. As in our case, the geometry was monitored systemmatically throughout the wear process using data logging facilities, a number of parameters were measured during the process. For their particular application they used a two-disk machine which although enabling varying degrees of sliding and rolling to be achieved, allowed the possibility of "avalanching" of the wear process to take place because the wear track did not continuously create new areas, as in the crossed-cylinders machine. Also they did not attempt to relocate the wear track exactly after each measurement which made the real comparison of parameters before, during, and after wear, very difficult especially as in most cases only two tracks were taken in order to get a statistical average. In the processing of the data also, there is evidence of false values due, as they say themselves, probably to improper alignment of the specimen during measurement of the surface. This produces obvious discrepancies in not just one of the measured parameters, namely, the power spectrum, but it throws some doubt on the rest of the measured parameters. Good pre-processing of the data such as that described in Chapter 3 would considerably alleviate this sort of problem.

Grieve, Kaliszer and Rowe (1969) attempt to monitor the surface geometry during the wear process. They use both relocation techniques and data logging techniques to record a three-dimensional contour of the specimen from parallel tracks taken with a Talysurf along the specimen. They use a technique for this based on that of Williamson (1968) although not so refined in the maintenance of both the dc level and the alignment of the traces. They use two machines, one a pin and ring machine and the other a two-ring machine. In neither of these is the same unambiguity of result achieved as with the crossed-cylinders machine. In the experiment they measure both friction and wear and in the digital analysis use three-point analysis taking no account of the problems that this and sampling inevitably introduce. Only the R_a and slope parameters are evaluated. Again, only more so, their results depend critically on the quality of the data they obtain from their apparatus. They presume heavily that not only is the surface aligned to the Talysurf apparatus when being measured but also that there is no curvature or error of form present in the waveform itself, both of these conditions being difficult to ensure in a practical situation.

A number of workers have been active in the analysis of the fluctuations in the frictional force, in particular Strang and Lewis (1949) and Rabinowicz (1951, 1954, 1956 and 1957). Strang and Lewis attempted to estimate the magnitude of the fluctuating component by measuring the sum of the vertical movements (ride) of the upper specimen in its movement over the lower at the same time as they measure the frictional force. No attempt was made to relate the instantaneous behaviour of the friction graph with that of the ride or that of the actual surface profile. They concluded that the percentage of the fluctuating component to the mean was a small percentage, being of the order of 5%. Rabinowicz (1954) has shown that the fluctuations of the friction graph do appear to follow a Gaussian law; as verified in our experiments. He does not relate this to the surface finish or the ride. Also Rabinowicz (1955) attempts to simulate the autocorrelation of a friction graph by taking various models of the shear force-displacement graph. He does not take into account any random element in the positioning of the contacts themselves. His practical results indicate that under the limited conditions of his experiments where both surfaces were equally rough the distance of independence corresponded to about twice the average contact size. This tends to verify our results,where the one body is smooth compared with the other and the independence distance is equal to the plastic zone size.

Restating the work described in this chapter, it will be seen that, despite its limitations, it appears to be the first attempt to measure in a single co-ordinated experiment, the friction, the ride, and the surface topography (both in its original form and as modified by rubbing). One major result of this work has been to show that the experimental evidence forces us to the conclusion that these tribological features can be regarded as stochastic processes and are capable of analysis in terms of the type of theory developed in Chapters 4 and 5. The important distinction is that, whereas the earlier work has as its starting point a strong theoretical basis, the present discussion is firmly founded upon experimental measurements.

The results obtained have illustrated that the concept of a surface in terms of its main structure as determined by its independence (correlation)length and RMS value is useful in the prediction of the run-in profile characteristics and in determining the effective filtering action of the wear process itself. Under the conditions of the experiments, the results have also pointed to a transfer function of the friction process. Further, these exercises have shown the importance of the light or no load contacting conditions in determining the run-in profile and have thrown light on those topographic parameters most likely to change throughout the wear process and also on those likely to remain constant; the one being important in monitoring the wear process, and the other for prediction of the run-in profile. More work can be done using the accurate relocation profilometry to investigate the flow of metal in the wear track. The accurate dc relocation provides a powerful means of contouring across the wear track whereas the high radial relocation accuracy in addition enables parameter usefulness to be assessed.

Synchronism of the ride and friction waveforms have enabled information to be obtained about the nature of Coulomb friction. What remains to be done is a complete tie up between the friction, ride and profile waveforms. This would enable quantitative information to be obtained concerning plasticity and elasticity and in particular the elastic recovery. One very interesting exercise would be to cross-correlate the friction, ride, and profile graphs; an experiment which would need a little more refined apparatus.

In summarising, the following comments can be made:

(a) When random surfaces are used, the tribological characteristics like ride and friction are also stochastic functions.

(b) The power and versatility of the crossed-cylinders machine together with relocation techniques and data logging facilities have been proved beyond doubt.

(c) Possibly the most important point is that the model proposed to describe the surface geometry not only enables some predictions to be made about the main tribological features, it also gives an insight into how these effects occur.

7. THE GENERATION OF RANDOM SURFACES

7.1 Introduction

In Chapter 4 the model used for the statistical description of a typical manufactured surface consists of a profile having a Gaussian distribution of ordinate heights together with an exponential autocorrelation function. One of the main reasons for the choice of these parameters was the fact that a significant proportion of surfaces which have been measured in the past and many others which have been deemed to be typical of common engineering practice in a recent OECD programme (von Weingraber 1969) have characteristics in reasonably close accord with the chosen model.

The types of surface which show a particular similarity with the model include those produced by manufacturing processes which involve some random element, for example, sand blasting or grinding. These techniques, particularly grinding, are the finishing process for a high proportion of surfaces used today; other basic cutting processes are more commonly used for stock removal rather than finishing.

In considering the characteristics of surfaces as derived from digital presentation of profiles it is necessary to consider the extent to which the pre-processing of the data prior to the analysis has affected the shape of the autocorrelation function. An autocorrelation function which might be exponential in shape can easily be changed to a second-order form similar to a damped exponential cosine, by the introduction of a low-cut filter; as mentioned in Chapter 3 such a filter has to be used to remove unwanted curvatures and slopes from the original data. In other words many autocorrelation functions which might appear to be second order could in fact be exponential. Indeed the exponential correlation function may occur much more commonly than is suggested by published data.

The widespread existence of surfaces having a height distribution which is close to Gaussian is even more clearly established (see for example, Pullen, Hunt and Williamson 1969). It therefore seems relevant to enquire whether the model adopted in Chapter 4 is to be expected when surfaces are generated by mechanical methods involving random processes. This question is, in turn, related to one of the central problems of the modern subject of surface typology discussed in Chapter 2; that is, whether it is valid and meaningful to classify surfaces by a specification of their height distribution and autocorrelation functions. (Unless stated the normalised autocorrelation will be considered).

To explore these questions it is necessary to develop a theory of the generation of surface profiles by random methods. To ensure its wide applicability, the method adopted here will be quite general. It will be assumed that the profile is produced by a series of unit events, each event possibly involving the removal, or movement of material in small quanta. The unit events will occur in a random fashion at points along the profile and the effect of changes in the nature of the unit event (including random variations in its shape and size) will be investigated. It is clear that such random elements of the process will occur in practical machining methods but the details of how this occurs in practice will not be considered at this stage. It is hoped that this approach will provide some physical insight into the relationship between the characteristics of the generated profile and the details of the mechanisms involved in its generation.

7.2 Simple simulation experiments

7.2.1 Method

The method adopted to explore the nature of the generated profile shape is by a simulation of the mechanics of the manufacturing process on a digital computer. As an example, in this technique the quanta representing the machining elements on a grinding wheel, the grits, are first specified as having a certain shape and size. The original profile is taken to be a horizontal straight line, represented in digital form, and extending over some 5 000 locations. In what follows, for the sake of simplicity, the machining elements will be called grits. In practice, the element could be a bead, a shot, or even an ion. In each computer "experiment" the shape of the unit events and their height distribution will be specified.

The position of the grit horizontally with respect to the surface is decided by a random number generator which gives a number between 1 and 5 000; this number is then taken as the location in the profile where the grit impinges. Another number representing the height of the grit is generated subject to rules

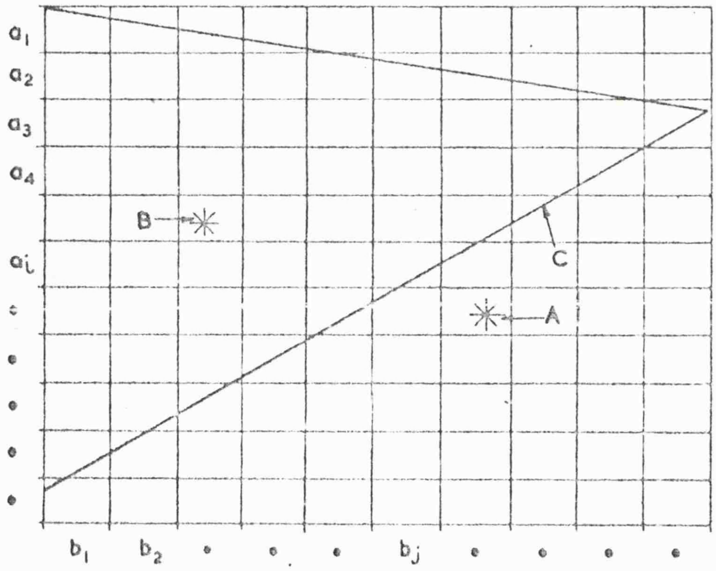
FIG. 7-1

GENERATION OF ANY HEIGHT DISTRIBUTION

A IS A NON-ACCEPTABLE POINT

B IS AN EXAMPLE OF AN ACCEPTABLE POINT

C IS A CURVE OF ANY REQUIRED DISTRIBUTION



explained later; this represents the height, relative to a zero taken as datum, at which the tip of the grit is presented to the surface. The shape and size of the grits are also decided, where appropriate, by another random number. If the grit so generated interacts with the profile at the appropriate location then a 'hit' is recorded; the height of the profile at that location is then modified to corresponding to this interaction. Locations immediately adjacent to it are then modified in accordance with the assumed shape, size and mechanism of interaction of the grit. After many such hits a profile has been built up which can then be analysed in much the same way as real surface profiles. This will be described in Section 7.2.3.

In this whole process a very important feature is the random number generator. The pseudo random number generator used here has the following formula. If RN_i is the new random number and RN_{i-1} the previous one then

$$RN_i = (\pi + RN_{i-1})^8 - \text{integer } (\pi + R_{i-1})^8 \qquad (7\text{-}1)$$

This yields a random number between 0 and 1 uniformly random to eight decimal places. The first four of these digits are used to define the horizontal location, the next two, the height at which the grit is presented and the last two the size of the grit. The shape of the grit is determined by the mode of operation of the computer program. This generator gives about ten thousand numbers before any danger of cycling occurs.

To generate a height distribution of grits which is other

FIG. 7-2

DISTRIBUTION OF POSSIBLE CUTTING POINTS.

LINE (a) IS ORIGINAL METAL.
LINE (b) IS MAXIMUM ALLOWABLE PENETRATION.

DISTRIBUTIONS.

(c) IS UNIFORMLY DISTRIBUTED FROM THE UPPER LEVEL (a) TO LOWER (b).
(d) IS LINEARLY DISTRIBUTED - CAN BE APPROXIMATELY EXPONENTIAL.
(e) IS TRIANGULARLY DISTRIBUTED - CAN BE APPROXIMATELY GAUSSIAN.

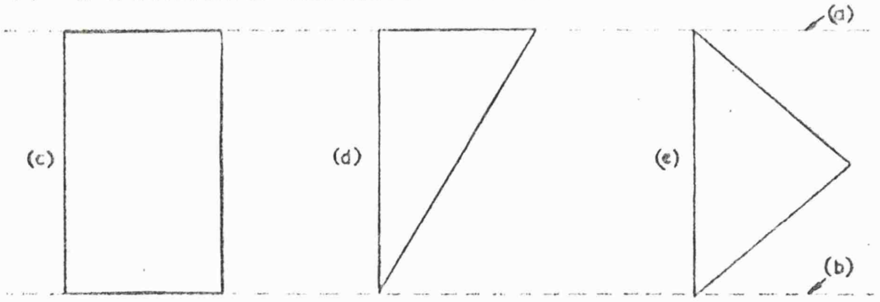
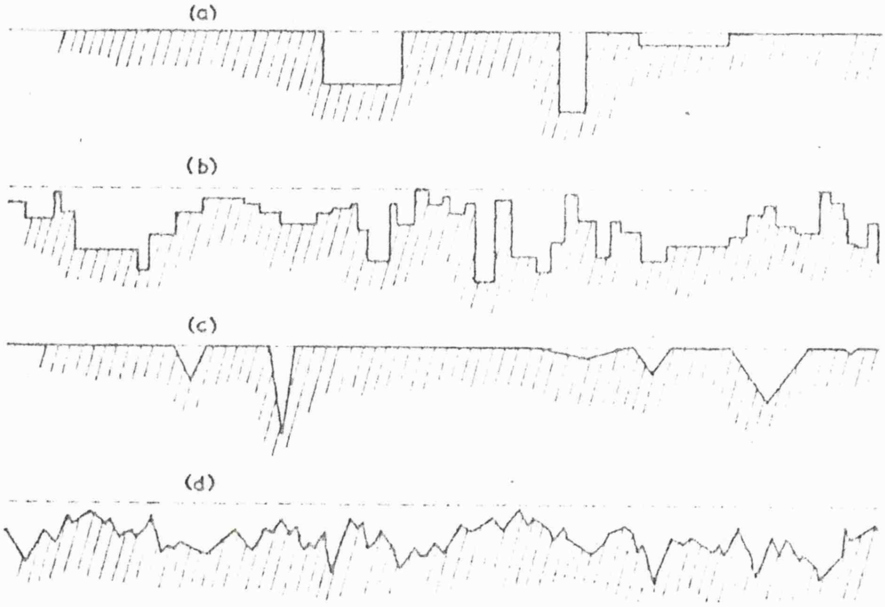


FIG. 7-3

GENERATION OF RANDOM PROFILE.

(a) FEW HITS - NO INTERACTION USING RANDOM WIDTH SQUARE GRIT.
(b) LARGE NO OF HITS USING RANDOM WIDTH SQUARE GRIT.
(c) FEW HITS - NO INTERACTION USING VARIABLE ANGLE TRIANGULAR GRIT.
(d) LARGE NO OF HITS USING VARIABLE ANGLE TRIANGULAR GRIT.



than rectangular is complicated. However, a method has been devised in which other desired height distribution shapes can reaily be produced. In this method the curve of the desired height probability density function is drawn relative to a two-dimensional vector matrix whose rows are a_i and columns b_j i.e. the equation of the curve in terms of a_i and b_j is found where, say, the height variable is a_i and the probability density b_j. This is shown for an almost symmetrical triangular distribution in figure 7-1. Any pair of co-ordinates represent a location in this matrix. If, when a pair of numbers has been generated the location lies outside the curve, say at position A, then the number is rejected, but on the other hand if it lies within the curve, say at B, then it is accepted and the value of a_i is used as the height at which the grit will be presented to the surface. In this way heights corresponding to any probability density are selected the correct number of times but in a random order. Using this technique, a height distribution having any shape within finite height limits can be generated. Some examples are shown in figure 7-2 (c), (d) and (e). These three height distributions have been used in these investigations. There are no reasons, other than those of operational convenience, why true truncated Gaussian or exponential distributions should not be employed instead of the approximate forms shown here which were used to represent them.

These three height distributions are of particular interest because of their use by research workers in this field. A Gaussian grit height distribution has been used by Baul (1967) in his work on simulated grinding. Orioda (1955), Yoshikawa and Sata (1968) and Yoshikawa and Peklenik (1968 and 1970) have used the linear distribution shown in figure 7-2, or similar shaped distributions. Yoshikawa and Peklenik (1970) use three different distributions corresponding to different dressing conditions of the grind wheel; of these three, two correspond to the linear and rectangular distributions which we use. Because of the difficulties involved in predicting the most significant distribution to use we rely, for our justification of the use of such height distributions, upon the practical considerations taken into account by these researchers.

The other consideration relevant to our investigations is the way in which the grits or events are distributed across the surface. In earlier work by other workers this distribution of grit positions is assumed to be uniform. What this implies and the tests used to ascertain the statistics of this distribution will be described in Section 7.3.2.

In our experiments two basic shapes of grit have been considered, the rectangular grit, and the triangular grit. The former is generated in two ways either having fixed width or alternatively a random width; which of these two modes is used depends on the path through the computer program. If the mode is fixed width it can be set by means of an input card. To illustrate how the rectangular grit operates on the surface consider one grit hitting the surface at a depth d below the surface level, where d refers to the depth of the centre of the grit. The program first positions this centre ordinate relative to the surface; it then removes metal on either side of this centre ordinate, at the same depth d, until the overall width of the grit impression agrees with that of the width of the grit previously determined. For fixed width mode this width is set at the beginning of the run, for random width mode the grit width is set after each operation by means of one of the digits generated by the random number generator. The rectangular grit is sometimes referred to as a square grit for simplicity. For simulated machining using triangular grits, four modes are possible, not including the ploughing mode which is described later in Section 7.4. The first is when the grit angle is random. Here, as in the square grit case, the depth (determined from the random number generator) to which the grit indents into the surface refers only to the centre ordinate of the grit; adjacent ordinates of the grit indent to a correspondingly smaller depth depending on the angle of the grit and the distance of the ordinate from the centre of the grit. In the random mode the random angle is determined by setting, from the random number generator, the width at the top of the grit. This, when taken with the depth of cut, provides a random angle. For the fixed mode, the width is restricted to one value, but this does not fix the angle but just the width at the top of the grit. Any degree of randomness of the width can be achieved by relaxing slightly the restriction on the values that the grit width can take. In another fixed mode the angle itself could be fixed completely or held between certain limits merely by selection from all the random possibilities generated. A final mode of operation for triangular grit was to suppress the one side of the grit completely thus in effect simulating a sawtooth grit.

Other facilities offered by the program (called GRIN and discussed in Appendix 2) include the ability to be able to simulate sparkout by dropping the datum from which the height distribution is measured at any stage during the "machining operation". It is also possible to operate in a mode in which the height distribution has no fixed datum but each event simply takes away from the existing surface the depth and width of grit that the random number generator has selected.

7.2.2 Development of profile waveform

In Section 7.2.1 the method in which one grit indents the surface has been discussed, together with various forms the shape of the grit might take (in these simple simulations). In this section we will discuss the characteristics of the surface profile which have resulted after a large number of operations. In this connection the term "operation" implies one change of the random number, it may or may not, imply a hit and a consequent change in the surface profile. Figure 7.3(a) shows the form that the profile has after only a relatively few hits, say 50, have been made using rectangular grit; the original height of the surface can be clearly seen. Figure 7.3(b) shows the profile that has developed after a large number of operations, again for rectangular grit. By comparison, figures 7.3(c) and 7.3(d) show how the profile has developed for the same number of operations using a triangular grit. It will be seen that in figure 7.3(b), and even more in figure 7.3(d), that the surface has taken on a noticeably random appearance. As will be shown later this random appearance does not depend on the assumed distribution of heights of the grits.

7.2.3 Characteristics of the generated profile - the amplitude distribution of the roughness

In this section the generated shape of the ordinate height distribution of the profile is considered in relation to the imposed height distribution of the grits. Although in principle any height distribution of grits could be investigated; accepting that in some cases the distribution would have to be somewhat truncated, only the three mentioned in Section 7.2.1, namely, rectangular, triangular and sawtooth, will be considered here.

Consider, to start with, the case when the grits are rectangular, of random width and of rectangular height distribution. At first, when only a few hits have been made, most of the original surface is left. Similarly, after a very large number of hits the profile again approaches a straight line corresponding to the depth of the deepest depth of removal allowed by the height distribution of grits. It was found that, in the case of the square grit, the "metal" is removed too fast for the roughness to be a significant part of the profile when compared with either the original or final straight line. Therefore an experiment which may be thought to be equivalent to plunge grinding was introduced, in

FIG. 7-4

AMPLITUDE DISTRIBUTIONS - RANDOM SQUARE GRIT

5000 OPERATIONS

(a) SAWTOOTH DISTRIBUTION (b) TRIANGULAR

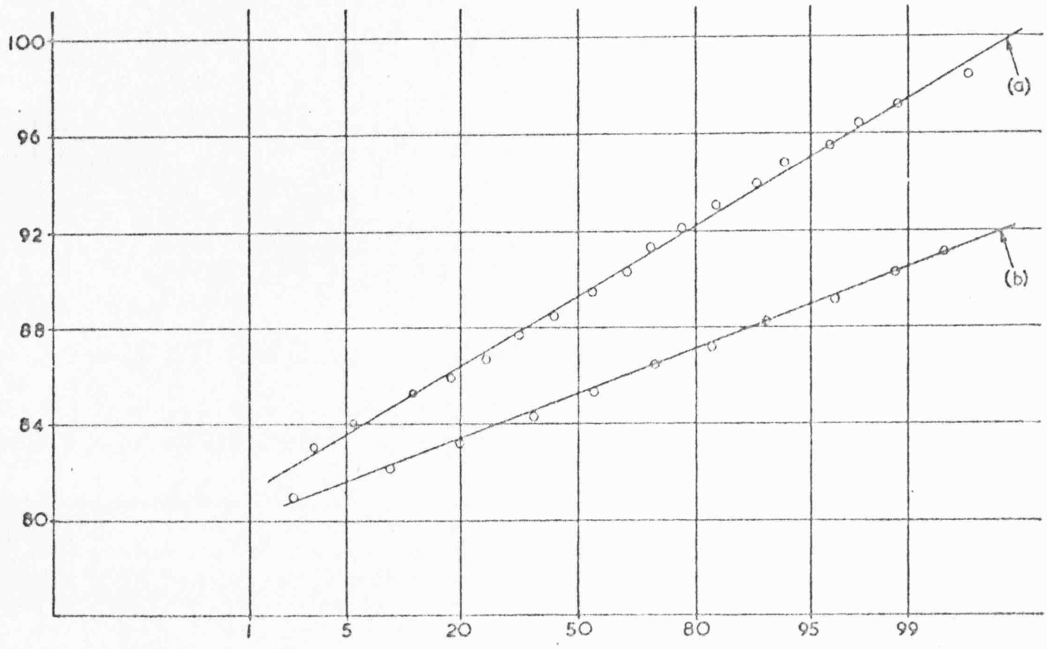


FIG. 7-5

AMPLITUDE DISTRIBUTIONS - RANDOM SQUARE GRIT

2500 OPERATIONS

(a) SAWTOOTH DISTRIBUTION (b) TRIANGULAR (c) RECTANGULAR

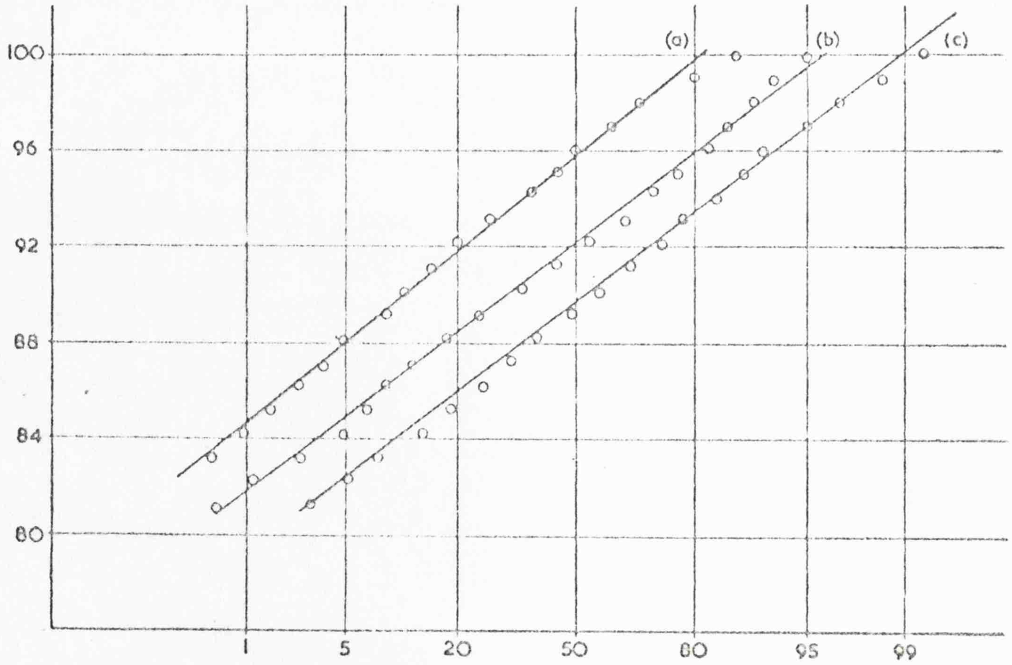
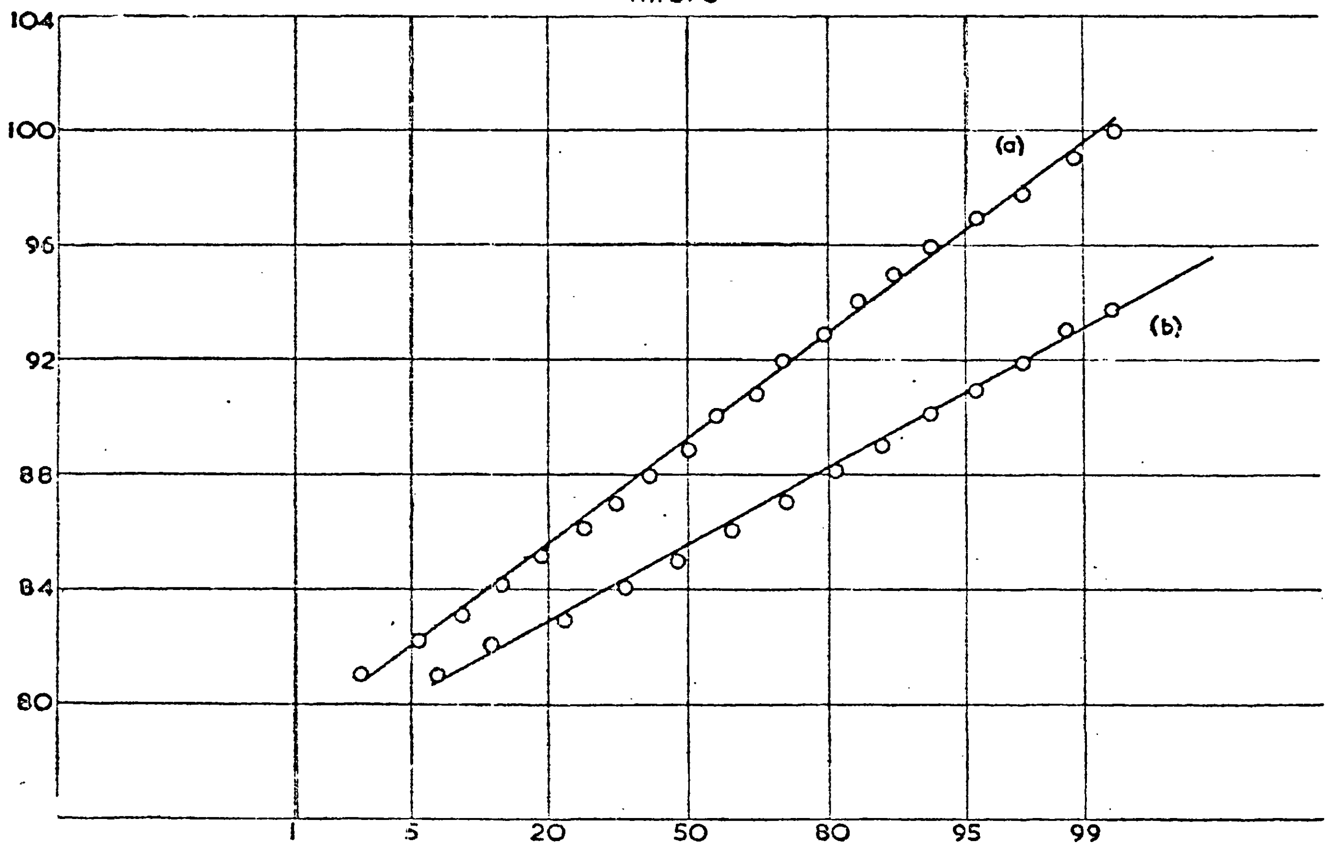


FIG.7-6

COMPARISON OF DISTRIBUTIONS OF (a) RANDOM ANGLE AND (b) FIXED ANGLE

1500 OPERATIONS

AV. OF 3



this only a limited number of operations or hits are allowed from any one datum. Under these conditions the roughness waveform exists. Figure 7-4 shows two typical distributions plotted on probability paper for the cases where the height distribution is triangular and sawtooth. The profile ordinate height distribution obtained in both cases are very close to Gaussian because the plots are straight lines on the probability paper. The same exercise has been performed with triangular grits both for the case of random angle and variable angle. In these cases the "metal" is not removed with quite the rapidity as when using rectangular grit with the rectangular height distribution; thus the results obtained for all grit height distributions can be displayed as shown in figure 7.5. Again, the height distributions are very close to Gaussian. This result has been found to be quite general. No matter what assumed height distribution of grits or whatever the shape and size of the grit it has been found that the ordinate height distribution of the roughness profile invariably has a Gaussian shape. (This distribution may be truncated at the top, after a few operations, or at the bottom, after very many operations, as explained above). Two further examples are shown in Figure 7.6.

Summarising the outcome of the simulated experiments over the range of conditions tried so far poses the following question. Is it true that the ordinate height distribution of a surface produced by a single manufacturing process having a decidedly random component is always Gaussian? The results would suggest that this is so.

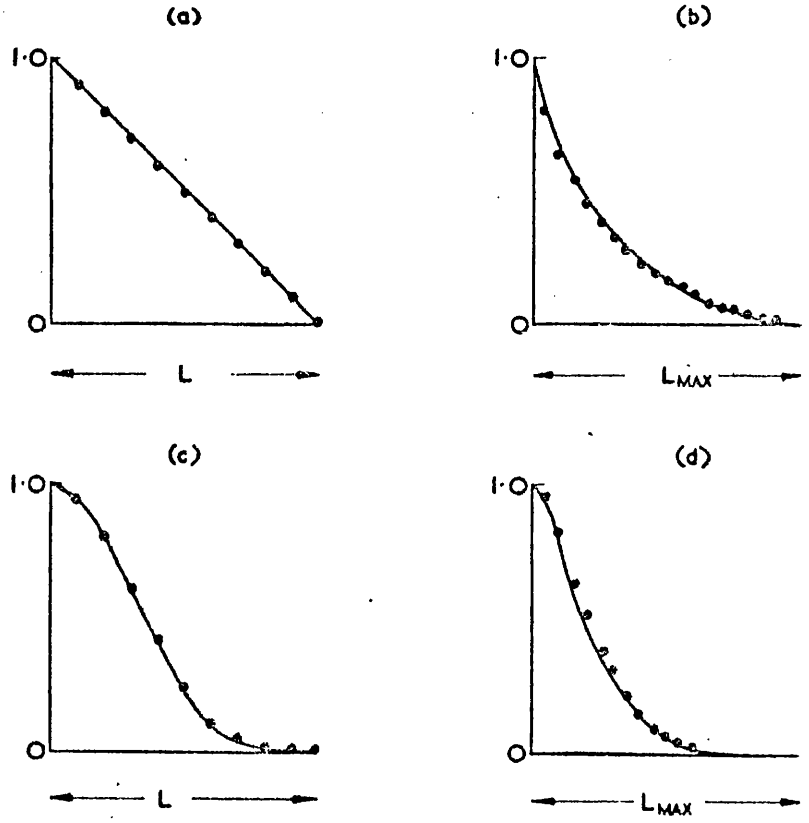
FIG.7-7

AUTOCORRELATION FUNCTION GRIT IMPRESSIONS WITH NO INTERACTION.

(a) FIXED WIDTH SQUARE GRIT. (b) RANDOM WIDTH SQUARE GRIT.

(c) FIXED WIDTH TRIANGULAR GRIT. (d) RANDOM WIDTH TRIANGULAR GRIT.

L IS FIXED WIDTH. L_{MAX} IS MAXIMUM WIDTH. DOTS ARE PRACTICAL POINTS.



7.2.4 Characteristics of the generated profile - the autocorrelation function

We turn our attention to the autocorrelation function of the profile produced by computer simulation using the techniques described above. In this work the same problems exist as have been described in discussing the generated height distribution. In deriving the autocorrelation function, it is even more important to ensure that all traces of the original straight line profile have been removed and also that the process has not proceeded too far and produced truncation of the height distribution at the lowest level because of the number of levels allowed in the array

In an attempt to provide some physical insight into our consideration of the shapes of autocorrelation functions it is instructive not only to consider that of the generated profile itself but also that of the individual grit impressions. Figure 7.7(a) shows the correlation function of a rectangular (or square) grit of fixed width and figure 7.7(b) shows the autocorrelation function for the square grits having random widths with a rectangular width probability density function (these results are proved in Section 7.3.2 below). Notice that in both these cases the correlation length (equivalent to $2.3\beta^*$ in Chapter 4) is determined in the first instance by the fixed width of the grit, and in the second by the maximum width of grit. In other words, for single grit impressions, without interaction between impressions, the length of the autocorrelation function is determined by the grit width (or strictly by the width of the

FIG. 7-8

EFFECT OF GRIT INTERACTION ON CORRELATION FUNCTION

ALL THE FOLLOWING ARE FOR SAWTOOTH DISTRIBUTION OF GRIT HEIGHTS

(a) REPRESENTS AUTOCORRELATION FUNCTION OF FIXED WIDTH SQUARE GRIT

-------- IS FOR NO INTERACTION

———— IS FOR INTERACTION PRODUCED AFTER 2500 OPERATIONS

○ ○ ○ ○ ○ IS FOR THEORETICAL VALUES

(b) REPRESENTS RANDOM WIDTH SQUARE GRIT

-------- IS FOR NO INTERACTION

———— IS FOR INTERACTION PRODUCED AFTER 2500 OPERATIONS

—·—·— IS FOR INTERACTION PRODUCED AFTER 5000 OPERATIONS

(c) REPRESENTS FIXED WIDTH TRIANGULAR GRIT

-------- IS FOR NO INTERACTION

———— IS FOR INTERACTION PRODUCED AFTER 2500 OPERATIONS

—·—·— IS FOR INTERACTION PRODUCED AFTER 5000 OPERATIONS

(d) REPRESENTS VARIABLE WIDTH TRIANGULAR GRIT

-------- IS FOR NO INTERACTION

———— IS FOR INTERACTION PRODUCED AFTER 2500 OPERATIONS

—·—·— IS FOR INTERACTION PRODUCED AFTER 5000 OPERATIONS

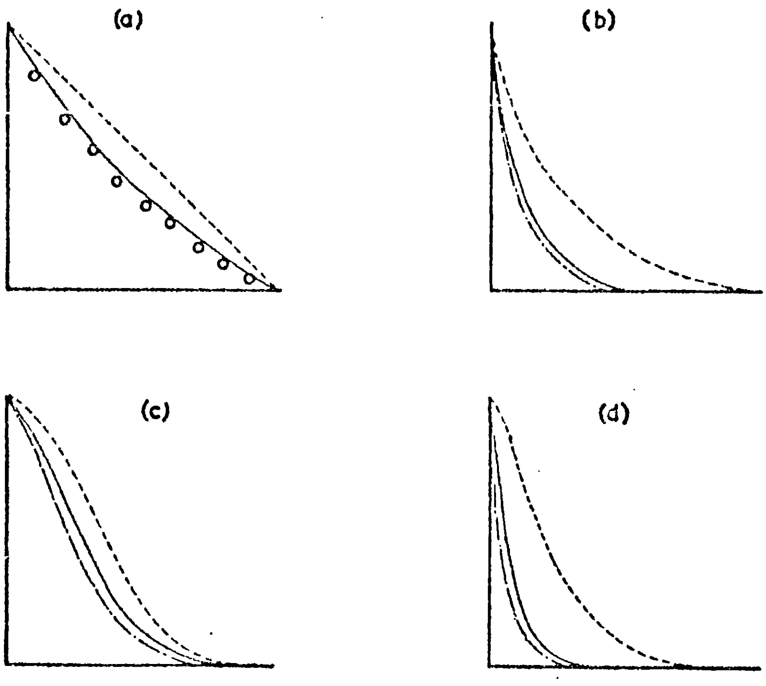
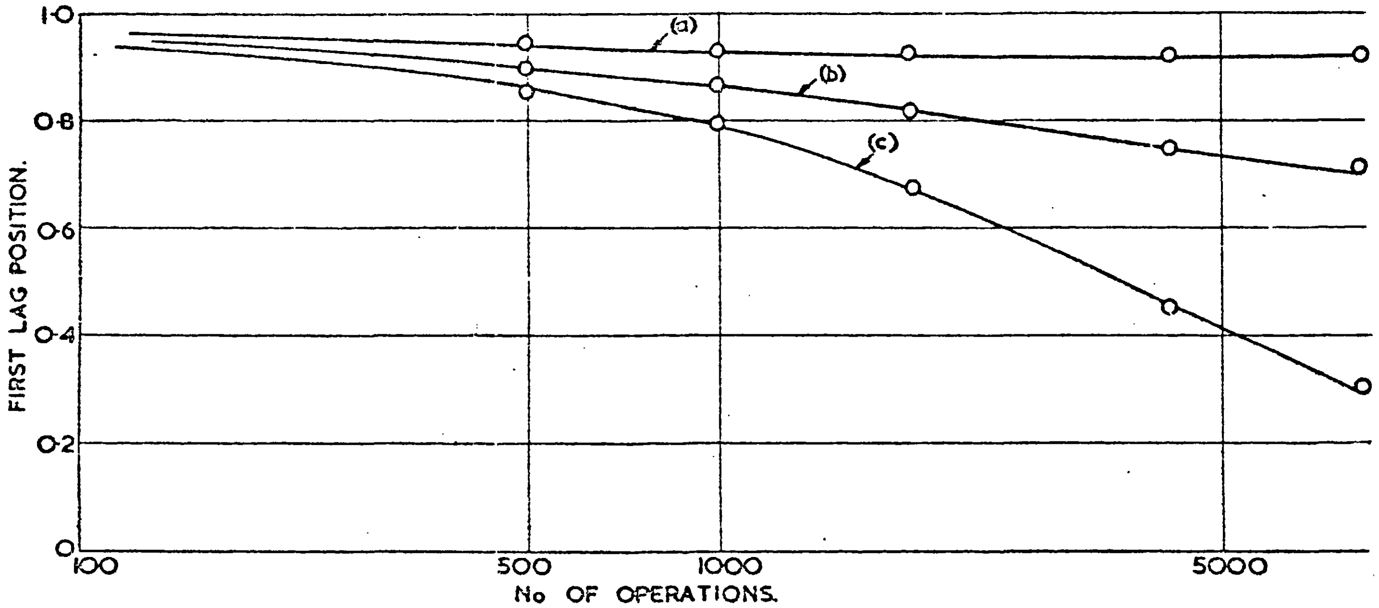


FIG 7-9

SHOWS RATE OF DROP OF FIRST LAG POSITION ON AUTOCORRELOGRAM AS A FUNCTION OF OPERATIONS A SAWTOOTH GRIT HEIGHT DISTRIBUTION IS USED

(a) IS FOR FIXED SQUARE GRIT (b) IS FOR VARIABLE SQUARE GRIT AND
(c) IS FOR RANDOM TRIANGULAR GRIT.



tip of the grit hitting the surface). It is difficult in the case of the random width grit to simulate the overall behaviour without running into problems of grit impression overlap. Figures 7.7(c) and 7.7(d) show the corresponding autocorrelation functions for triangular grits using both fixed angle (figure 7.7(c)) and random angle (figure 7-7(d)). Again, as for the square grit, the significant extent of the autocorrelation function is determined by the grit size.

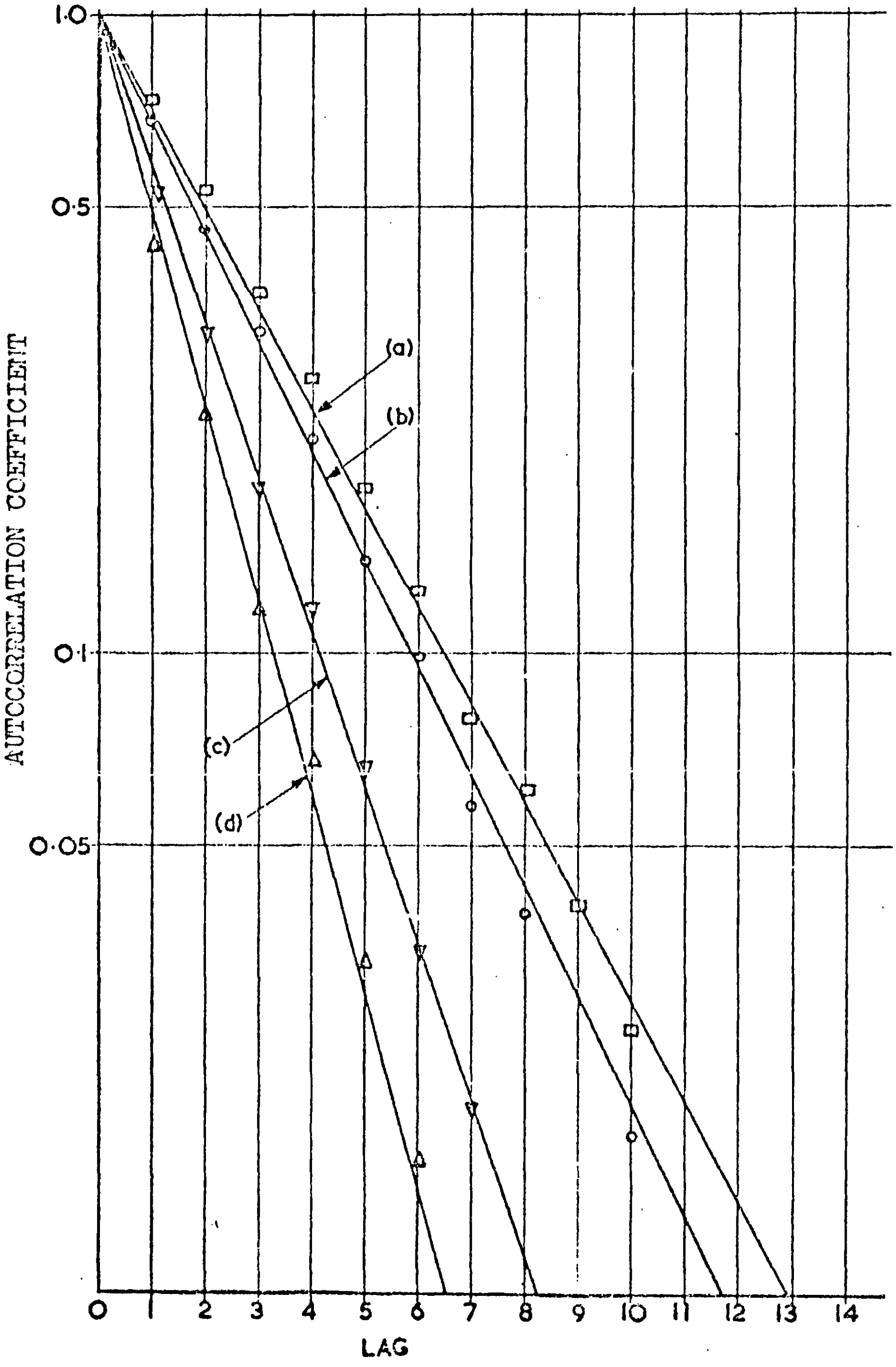
Consider now the situation in which the density of grit hits is large enough to ensure that a large proportion of overlaps occur. Figure 7-8 shows a series of results obtained from computer simulation, the correlation function for individual grits being shown as a broken line, in each example, for comparison.

Certain broad conclusions can be drawn from these results. First, the correlation function, originally considered as that for the individual grit impressions, is modified by the existence of interactions between grit impressions; the effect of this interaction is, in general, to shorten the length of the correlation function and to change its shape. Second, as the interaction proceeds, with an increase in the number of operations, these influences become more marked. The influence of the number of operations upon the slope of the autocorrelation function near the origin is shown in figure 7-9. Here, what is actually shown is the value of the autocorrelation function one ordinate spacing from the origin. Because the surface is stored as an array in the computer the first ordinate position refers to a shift of just one location,

FIG. 7-10

EXPONENTIAL PLOT OF GRITS

(a) VARIABLE SQUARE GRIT TRIANGULAR GRIT DISTRIBUTION
(b) VARIABLE SQUARE GRIT SAWTOOTH GRIT DISTRIBUTION
(c) VARIABLE TRIANGULAR GRIT DISTRIBUTION AS (a)
(d) VARIABLE TRIANGULAR GRIT DISTRIBUTION AS (b)



the minimum change possible. This gives an indication of slope at the origin and can be used as an estimate of the distance over which the correlation function is large valued. Third, figure 7-8 suggests that when both random interaction and random shape of grits are operative the autocorrelation function of the generated profile tends towards an exponential shape. To examine this point in more detail some autocorrelation functions of profiles generated after 5 000 operations, with random grit shape and with interaction of events, have been plotted on a log-linear scale in figure 7-10. The plots are very close to linear. It would appear, therefore, that the introduction of a larger element of randomness into the generation process produces profiles whose correlation functions are, indeed, close to the exponential form. The relative influence of the grit height distribution and the grit shape upon the length of the autocorrelation function will also be observed in figure 7-10.

7.2.5 Discussion of simple simulation experiments

Summarising what has been said above in Sections 7.2.3 and 7.2.4 the following conclusions may be reached:

1. All the grit shapes and height distributions used in these simulations tend to produce Gaussian height distributions of the generated profile.

2. The length of the autocorrelation function is not only determined by the size of the grit but also the degree of interaction between grit impressions which tend to reduce it.

3. When the degree of randomness in the grit sizes, shapes, and interactions becomes high the autocorrelation function of the generated profile tends to become more nearly exponential in its shape.

These conclusions suggest a need for a more mathematical approach to the theory of the characteristics of the profile generated by random processes. This will be provided below in Section 7.3. The influence of other factors upon the characteristics of the generated profile have also been explored in computer simulation experiments. These experiments and the results will be discussed in Section 7.4.

7.3 Theory

7.3.1 Height distribution

In this section some consideration will be given to the height distribution of a profile of a surface produced by random manufacturing process. From the point of view of the height distribution these random processes fall roughly into two categories; those like shot peening, flame spray deposition etc. in which particles impinge onto the surfaces. Most of the roughness in these cases is produced through random plastic flow or random deposition. Others (e.g. grinding, lapping) involve a degree of metal removal and in addition the elementary cutting particles are firmly bonded or embedded into a toolpiece.

In both of these types of manufacturing process, and perhaps others, the conditions are such that it is not unrealistic to suggest that the shape of the ordinate height distribution of the profile could be a direct result of the Central Limit Theorem, often used in statistics to explain the preponderance of Gaussian distributions arising in nature. This says, in effect, that a Gaussian distribution will occur quite generally as a result of a large number of independent random variables, of the same order of magnitude, acting together.

To be more specific. If there are N mutually independent random variables whose individual distributions can be different, and even unspecified, then the distribution of the sum of these variables tends to be Gaussian as N becomes large (see Cramer 1946). The Central Limit Theorem does not only apply to probabilistic considerations but is also a property of repeated convolutions and is used in many fields, for instance, in electrical network theory. As an example, the impulse response of a large number of cascaded filters is Gaussian whatever the shapes of the individual filter responses.

In the manufacturing processes involving the random addition, removal, or impinging of metal particles, the roughness waveform that eventually results is the cumulative effect of many local plastic flows resulting from separate hits. The complexity of such a situation is considerable after only a few hits. Because the roughness is effectively produced by many independent particle hits, or removals, then the amplitude distribution could, to some extent, be justifiably expected to be Gaussian no matter how each of the individual hits has affected the surface profile. This type of process satisfies a basic philosophy of the Central Limit Theorem which is that the process variables should be additive; this is done because the roughness profile is generated as a result of the total history of the process extending well into the immediate past.

The situation in some of the other processes is not so straightforward because each cutting grit height is fixed relative to some plane determined, for example, by the grind wheel or the lap. Consequently it might be argued that tne surface roughness produced by such a process would be automatically the inverted distribution of the cutting elements on the wheel or lap. This is not necessarily so. The roughness height probability is determined to a large extent by the random element in the horizontal positions of the cutting elements. This is especially true the more random is the shape of the cutting element. The importance of random positioning is met with in communication theory where it can be shown, for instance, that the height distribution of the sum of a number of sine waves having random phase approaches a Gaussian distribution as the number of components increases (Panter 1965).

The condition that the cutting elements in the process are independent of each other is substantially true even in the case of a grindwheel because it is usually only the outermost elements that actually contribute to the cutting action and for levels far away from the mean level they can be considered to be more or less independent. Under some conditions the requirement for independence as a pre-requisite of the Central Limit Theorem can be relaxed (Cramer 1946).

It is not so easy to argue the additive property of these kinds of process but it would certainly seem that the roughness in any area on the surface in practical situations would, in general, not be dependent on simply the last grit but it would be dependent on all the past history of grits hitting in the neighbourhood; the final result being determined in a complex way. In all practical cases (e.g. grinding) there is a certain amount of metal flow involved in a hit as well as metal removal, this makes these processes more akin to the bead blasted or shot peened type of process, at least in terms of the geometrical generation of the surface. However, from the results of figures 7-4, 5, and 6 it would appear that the Central Limit Theorex does hold even for the metal removal processes only, for quite a general distribution of grit heights.

7.3.2 Autocorrelation function

We turn now to a consideration of the autocorrelation function of the generated profile. Because we are usually concerned with the shape the normalised autocorrelation function will be frequently considered (Davenport and Root 1958). It is to be expected that a number of factors could enter into the problem; these might include:

(a) the randomness of the process (i.e. how the grits are distributed and,

(b) the shape of the grits as well as the way in which the grit reacts with the surface.

In order to be able to build up a picture of how these characteristics of the profile affect the profile autocorrelation function we will first consider each one in turn and will then bring them together in a more comprehensive picture.

In order to be able to consider the development of the autocorrelation function we must first consider the statistical character of the spatial distribution of the individual events; taking the co-ordinates of the centre of the cutting elements into account only and not their shape or size. To get an idea of the basic statistics of the practical situation it is informative to approach the problem from a Bernoulli trial theory. Thus, if in a large interval 0 to L, N events have occurred (where N is large) then the probability of finding K and only K events in an interval $\ell << L$ is given according to Bernoulli by

$$\text{Prob (No in } \ell = K) = {}^{N}C_{K}\, p^{K}\, q^{N-K} = \frac{N!}{K!(N-K)!}\, p^{K}\, q^{N-K} \tag{7-1}$$

where $p = \ell/L$ and $q = (1-p)$; p is the probability of finding a specific point in ℓ.

If $N >> 1$ and $\ell/L << 1$ then equation (7-1) reduces to Poisson's theorem in which

$$\text{Prob (No in } \ell = K) = \exp(-N\ell/L) \; . \; \frac{(N\ell/L)^K}{K!} \qquad (7\text{-}2)$$

which when writing $\lambda = N/L$ becomes

$$\exp(-\lambda\ell) \; . \; \frac{(\lambda\ell)^K}{K!} \qquad (7\text{-}3)$$

It is well-known that if $\ell = \ell_1 - \ell_2$ and λ is not constant but a function of x i.e. $\lambda(x)$ then

$$\text{Prob (No in } \ell = K) = \exp\left(-\int_{\ell_1}^{\ell_2} \lambda(x)dx\right) \; . \; \frac{\left(\int_{\ell_1}^{\ell_2} \lambda(x)dx\right)^K}{K!} \qquad (7\text{-}4)$$

Consequently for an interval 0 to β this would become

$$\exp(-\lambda\beta) \; . \; \frac{(\lambda\beta)^K}{K!} \qquad (7\text{-}5)$$

for the special cases which we shall require shortly (K = 1), the probability that only one event occurs in the interval between 0 and β is

$$\exp(-\lambda\beta) \; . \; \lambda\beta \qquad (7\text{-}6)$$

and also for the case where no events (K = 0) lie in the same interval then equation (7-5) reduces to

$$\exp(-\lambda\beta) \qquad (7\text{-}7)$$

For events to be classified as Poissonian it is usually necessary for some requirements to be met. These are that the probability of more than one event happening in a small interval $\delta \ell$ is zero (remembering here that only the co-ordinate of the event position is being considered), that each event is of an independent character (grits are independent) and that what happens in one interval is independent of what happens in another.

In the context of this chapter one thing to bear in mind is that it is in space only that events are considered. We are looking at the spatial consequences of what has been happening in time. It will be obvious in machine processes like shot blasting etc. that the position of the individual grits in space must be independent of each other. In the case of grinding or polishing the situation is not so simple because the grits are all bonded together in the wheel. However, as in the justification for the Central Limit Theorem, mainly the tips of the higher grits are involved and they can reasonably be considered independent.

In order to ascertain whether grit peaks on a typical grindwheel for instance could be considered Poissonian, a standard test (Parzen 1962) was carried out on a number of profiles of a grinding wheel. A blunt stylus was used to pick out only the higher grits; then the distance of each peak, so revealed, from an arbitrary starting point was measured. Let these distances be designated U_1, U_2, ...U_N for N peaks. If the events have occurred in accordance with a Poisson point process these random variables U_i will be independent and uniformly distributed over the length of the profile.

It is possible to make a test for both uniformity and independence by observing that if U_i are independent then using the Central Limit Theorem for moderately large values of N these values will be arranged according to a Gaussian distribution hence their sum

$$S_N = \sum_{i=1}^{N} U_i$$

will have a Gaussian distribution with a mean of NL/2 and a variance given by

$$\mathrm{Var}\ (S_N) = \mathrm{NVar}\ (U_i) = \frac{NL^2}{12} \qquad (7\text{-}8)$$

yielding a standard deviation σ.

The test is that if S_N lies between $\overline{U} - 2\sigma$ and $\overline{U} + 2\sigma$ then there is a 95% probability that the test would indicate correctly that the process is Poissonian.

A typical trace showed $S_N = 884$, $\overline{U} - 2\sigma = 757$, and $\overline{U} + 2\sigma = 1011$, thus proving subject only to the confidence in the test that the high peaks can be considered to be Poissonian. All traces examined showed similar results. In fact doing similar tests on the profiles produced from these processes also gave the same conclusion i.e. the profile peak positions also appear to obey a Poissonian distribution. This is considered later. Hence as far as the positions of events are concerned the final profile can be

FIG. 7-11

MECHANISM OF AUTOCORRELATION FUNCTION DEVELOPMENT IN TERMS OF GRIT IMPRESSION

(a) POISSON IMPULSE TRAIN REPRESENTING CO-ORDINATES OF GRIT CENTRES - SHOWN DOTTED

(b) CONDITION WHERE $L/2 > t$

(c) CONDITION WHERE $L/2 < t < L$

(d) CONDITION WHERE $L < t$

(e) REPRESENTS GRIT

(f) REPRESENTS VIRGIN SURFACE

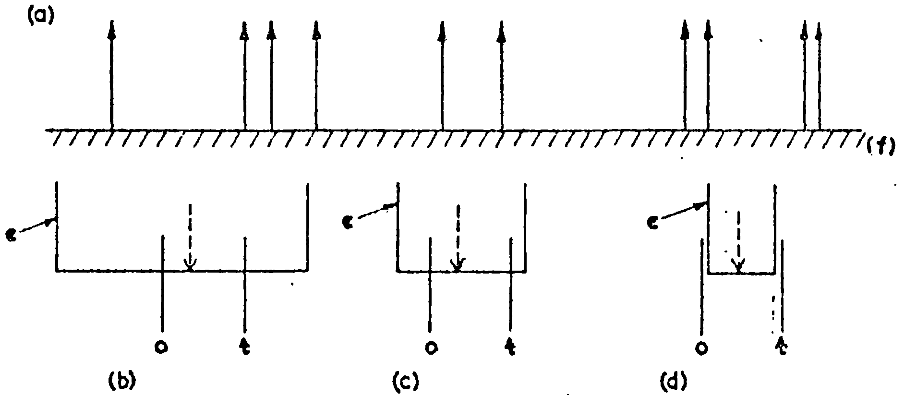
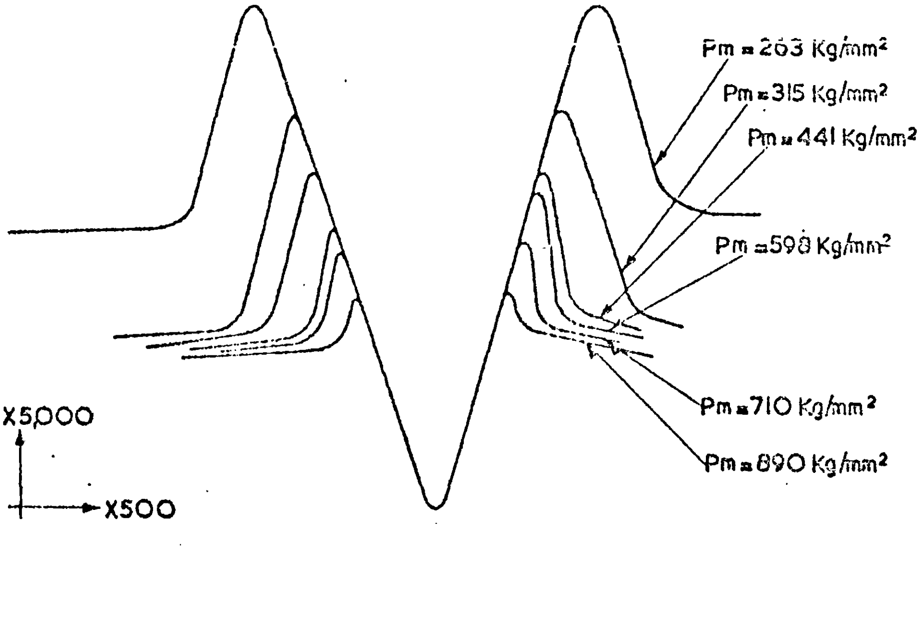


FIG. 7-12

SHOWING COMPOSITE PICTURE OF RELATION BETWEEN PILE UP AND VICKERS HARDNESS FOR 1% CARBON STEEL (AFTER BUTTERY)



considered to be made up of a succession of impulses in space, each representing the co-ordinate of a central position of a grit. Thus, one ingredient likely to be of importance in determining the autocorrelation function of the profile resulting from a random manufacturing process, namely the positions of cutting elements appears to be representable by a Poisson Impulse Train (figure 7-11(a)).

The next stage in the development of an understanding of the form that an autocorrelation function might take is to consider the autocorrelation function of a single event, i.e. the impression left by a single grit on the virgin surface. Consider first, for simplicity, the case of a square grit impression of known length L. The general autocorrelation function of a single impression, which is an aperiodic function, is given by

$$\int_0^{L-\beta} y_1(x)\ y_1(x+\beta)\ dx \quad (\beta<L) \qquad (7\text{-}9)$$

where $y_1(x)$ is the function denoting the grit impression shape. Note that it is valid to compare the analysis of a single grit as in equation (7-9) to the results obtained by computer from a consideration of a single grit impression in the whole length of a virgin surface providing that the width of the impression is small compared to the length of surface. For a single square grit impression, equation (7-9) becomes

$$\sigma^2 \quad (1 - \frac{|\beta|}{L}) \qquad (7\text{-}10)$$

where y is the depth of the square impression, which when normalised by σ^2 and calling $\beta/L = \overline{\beta}$ yields

$$C(\overline{\beta}) = (1 - |\overline{\beta}|) \qquad (7\text{-}11)$$

If, instead of a single impression of fixed width L a number of separated impressions of uniformly random width, and maximum width L_{max} are considered, then the autocorrelation function becomes

$$C(\overline{\beta}) = 1 - |\overline{\beta}| \, (1 + \ell n \, (1/ \, |\overline{\beta}|) \qquad (7\text{-}12)$$

$$\sim \exp \, (-|\overline{\beta}| \, / \, e)$$

In general if $C(\beta,L)$ is the autocorrelation function associated with an independent grit impression and $f(L)$ is the probability density of a grit having a maximum size related to L then a useful expression is

$$C(\beta) = \int_{\beta}^{L_{max}} C(\beta,L) \; f \; (L) \; dL \qquad (7\text{-}13)$$

where $C(\beta,L) = 0$ for $L < \beta$.

In the case considered above, when $f(L)$ is constant this yields equation (7-12) for a random square (or strictly speaking rectangular) grit.

In figure 7-7 some typical autocorrelation functions for single grit impressions are plotted together with computed values.

In the case of random widths obviously more than just one grit impression had to be put into the surface in the computer simulation but care was taken to ensure that none overlapped.

For a triangular grit having a constant width at the top of the impression the normalised autocorrelation function is given by

$$C(\overline{\beta}) = 1 - 6|\overline{\beta}|^2 + 6|\overline{\beta}|^3 \qquad \text{for } |\overline{\beta}| < 0.5$$

and $$C(\overline{\beta}) = 2 - 6|\overline{\beta}| + 6|\overline{\beta}|^2 - 2|\overline{\beta}|^3 \qquad \text{for } 0.5 < |\overline{\beta}| < 1.0 \qquad (7\text{-}14)$$

It may be noted here that a plot of equation (7-14) in figure 7-7(c) has a shape significantly different from that derived for a square grit (figure 7-7(a)). It is now very nearly Gaussian in shape and this is due to the fact that the convolution of a triangle is very close to Gaussian, in fact for the case under discussion

$$C(\beta) \simeq \exp\left(-5.55\ \beta^2/L^2\right) \qquad (7\text{-}15)$$

from which the autocorrelation function for triangular grit impressions having uniformly random widths at their base is given by

$$C(\beta) = \int_{\beta}^{L_{max}} \exp\left(-5.55\ \beta^2/L^2\right)\, dL \qquad (7\text{-}16)$$

which is shown in figure 7-7(d).

Thus in all these examples, where no interaction of grit indentations is assumed, the extent of the correlation function over which it has a significant value is determined by the assumed size of the unit event. Under conditions of independent events, without interactions, the correlation length (the separation of points on the surface which are effectively independent) is determined mainly by the size of the longest unit event; similarly under these same conditions, the shape of the correlation function is determined by the shape of the grit indentation.

However, if the number of grits or particles hitting the surface is increased such that interactions between grit indentations or impressions occur then the situation usually becomes more complex. Different effects can occur depending on the type of random process. Before considering processes like grinding, grit blasting etc. we will take a look at the effect of interaction in cases such as flame spray deposition or spark erosion. In these processes when there is an interaction between two unit events in the region of the overlap there is an addition (or in the case of erosion a subtraction) of the impressions. If one supposes that each metal globule (in the case of flame deposition) which is deposited on the surface has the same size and making the assumption that they are deposited in such a well-behaved manner that the superposition theory applies, this enables the mechanism of the process to be investigated simply. Under these circumstances the process can be considered to be the convolution of a Poisson impulse train (representing the positions at which the globules hit) with the shape of an individually deposited blob (or hole in spark erosion), i.e.

$$y(x) \quad = \quad Z(x) * h(x) \qquad (7\text{-}17)$$

where Z(x) is the impulse train and h(x) is the shape of the unit metallic deposit.

Using a result obtained in electrical theory (Papoulis 1965) enables the autocorrelation function to be obtained. It is given by C(β) where

$$C(\beta) \quad = \quad \overline{\lambda}^2 \int_{-\infty}^{\infty} h(\alpha) \;.\; h(\beta+\alpha) \; d\alpha \qquad (7\text{-}18)$$

which turns out to be the same as that obtained for a single unit event. In other words, given these circumstances, the correlation function would not degenerate with superposition of the unit events. Obviously this is only an approximate picture, but it does give an indication of under what conditions in the mechanism of the manufacturing process the autocorrelation function could be expected to remain independent of unit event interaction. Incidentally, it can be shown that if λ, the density of the Poisson impulse train, is high and considerable overlapping occurs, then in this electrical filter model for spark erosion the height distribution of the profile waveform tends to a Gaussian shape (Papoulis 1965) which is a direct result of the Central Limit Theorem. This illustrates that in this sort of process the Gaussian height distribution of the profile is a natural result.

An important distinction needs to be drawn at this stage. In the last paragraph it has been shown that when a series of events are superposed, without their modification in any way, then the correlation function of the generated profile is the unmodified correlation function of the unit events. However, in the simulation experiments described in Section 7.2 interaction occurred. By interaction we mean that the consequences of a selected event is determined, in part, by the state of the profile at the location in question. We consider below the consequences of this interaction.

To get an idea of how the autocorrelation is changed from that when the unit events are non-interacting consider figure 7-11(a) which shows a Poisson train of locations along a surface where the centres of the unit events making up the final profile are located. When working out an autocorrelation function of the profile for a lag β it is necessary to evaluate the ensemble average. product of two ordinates separated by β. Some of these products will be from situations where both ordinates are within one unit event, for some other products perhaps one ordinate lies in one unit event and the other ordinate lies in the adjacent unit event. (It must be emphasised here that by the very nature of the mechanism of the process and the interaction very few of the unit events will be whole; the very fact that interaction between them has taken place ensures this). Similarly, in other products more than two unit events or remnants of unit events may separate the ordinates.

The evaluation of the autocorrelation function may be summed up in terms of conditional expectation. Thus

$$E(\gamma) = E_{\zeta}(E(\gamma/\zeta)) \tag{7-19}$$

$$= \sum_{\substack{\text{over all}\\ \zeta}} E(\gamma/\zeta=\zeta)\ f(\zeta) \tag{7-20}$$

In equation (7-20) E(γ) is the expected value of situation γ and ζ is a conditional subset of γ.

In terms of the autocorrelation function

$$\begin{aligned} C(\beta) &= E(y(x)\ .\ y(x+\beta)) \\ &= E\left[y(x).y(x+\beta) \Big/ \begin{matrix}\text{both in}\\ \text{single event}\end{matrix}\right] x \left[\text{probability of single event}\right] \\ + \quad & E\left[y(x).y(x+\beta) \Big/ \begin{matrix}\text{ordinates in}\\ \text{two events}\end{matrix}\right] x \left[\text{probability of two events}\right] \\ &\cdot \\ &\cdot \\ &\cdot \\ \text{etc} \end{aligned} \tag{7-21}$$

where each line of equation (7-21) is strictly an ensemble average.

The essential feature about this equation is that E(y(x).y(x+β)) is zero except for the first case when the two ordinates are contained within a single unit event. This is because the unit events are unrelated so an ordinate from one unit event will, on average, not be correlated with an ordinate from another with the resulting effect of making the product zero. Thus E(y(x).y(x+β)) = E(y(x)).E(y(x+β)) = 0 for all but single unit event cases, assuming E(y(x)) = 0. The relationship between y(x) and y(x+β) when both are within one event depends upon the degree of randomness in the unit event itself; this will be discussed later. However, in order to see how this principle works out consider the simple case of a fixed width square grit unit event. Our problem is that of deciding the conditions under which two ordinates, separated by β, lie within the same unit event (grit impression on the surface) the actual value of $E(y(x).\ y(x+\beta))$ for a unit event is not difficult to get because for square grit $y(x)$ will be equal to $y(x+\beta)$ and the variance of the product will be the same as for the profile as a whole (assumed to be unity) hence the real problem is finding the probabilities of occurrance.

Consider figure 7-11. In order that the grit covers the interval 0 to β it must have a centre between $\beta - L/2$, and $L/2$, a total range of $L-\beta$. Using Poisson statistics whose density of occurrances if λ_p the probability of this happening is given by

$$\exp\ (-\lambda_p(L-\beta))\ \lambda_p\ (L-\beta) \qquad (7-22)$$

In addition no unit event centre must occur to cover up the origin but not β i.e. there must be no event centre between $\beta - L/2$ and $-L/2$ a range of β. The probability of this non-event is

$$\exp\ (-\lambda_p\beta) \qquad (7-23)$$

Similarly no second event must occur to include β but not the origin. The probability of this non-event is also $\exp\ (-\lambda_p\beta)$.

Hence the total probability of the two ordinates lying within one unit event is f(1) given by

$$
\begin{aligned}
f(1) &= \exp\,(-\lambda_p(L-\beta))\ \lambda_p\ (L-\beta)\ .\ \exp\,(-\lambda_p\beta)\ .\ \exp\,(-\lambda_p\beta) \\
&= \exp\,(-\lambda_p L)\ .\ \exp\,(-\lambda_p\beta)\ \ \lambda_p(L-\beta)
\end{aligned}
\tag{7-24}
$$

Hence C(β) may be written

$$C(\beta) = \text{constant}\ .\ \exp\,(-\lambda_p|\beta|)\ (1-|\beta|/L) \tag{7-25}$$

where the constant depends on λ_p and L.

Equation (7-25) is an interesting result because $(1-|\beta|/L)$ represents the autocorrelation function for the fixed width square grit without interaction. It appears that in this case the effect of interaction is to multiply the single grit result by an exponential term. That this equation has some basis of truth can be seen in figure 7-8(a), which shows the simulated result as well as the analytic values. In fact the treatment given above is a little too restrictive because the centres of other unit events contributing to the profile could lie in the forbidden regions. They would, however, interact if their level was lower than the unit event in the interval 0 to β. From this equation (7-25) and also equation (7-12) it seems as though increasing the random element either in positioning of the events or in their shape tends to make the autocorrelation function more like exponential in shape.

A further point indicates that this may well be so, because consider the case when the square grit has such random width and is so interacting that identity with a fixed unit event is blurred it then becomes more meaningful to consider the square edges in the resultant profile as a Poisson process. Using the test already described the relevant inequality obtained from a typical profile chart was $10540 < S_N = 13128 < 15140$ which means that one could accept the statistical hypothesis that the observed results were of a Poisson type. Under these circumstances equation (7-21) can be re-written in terms of number of edges rather than number of unit events; the criterion for two ordinates to be within one grit impression then becomes equivalent to them lying within no change of the profile level. Hence if the signal variance is unity and the Poisson density of edges λ_s then $C(\beta)$ becomes, using Poisson statistics

$$\begin{aligned} E(y(x).y(x+\beta)) &= \exp(-\lambda_s \beta) \; x\left(E(y(x).y(x+\beta)/\text{no change in level}\right) \\ &+ \lambda_s \beta \exp(-\lambda_s \beta) \; x\left(E(y(x).y(x+\beta)/\text{one change in level}\right) \\ &\;\vdots \\ &\text{etc.} \end{aligned} \qquad (7\text{-}26)$$

which reduces when normalised to

$$C(\beta) = \exp(-\lambda_s|\beta|) \qquad (7\text{-}27)$$

because the only contribution to the product sum comes from the first term, in all the other cases the levels are independent and consequently the product sum is zero. Notice that the result of equation (7-27) is different from that obtained for a similar sort of waveform, the random telegraphic signal, this is because of the multiplicity of levels in our profile.

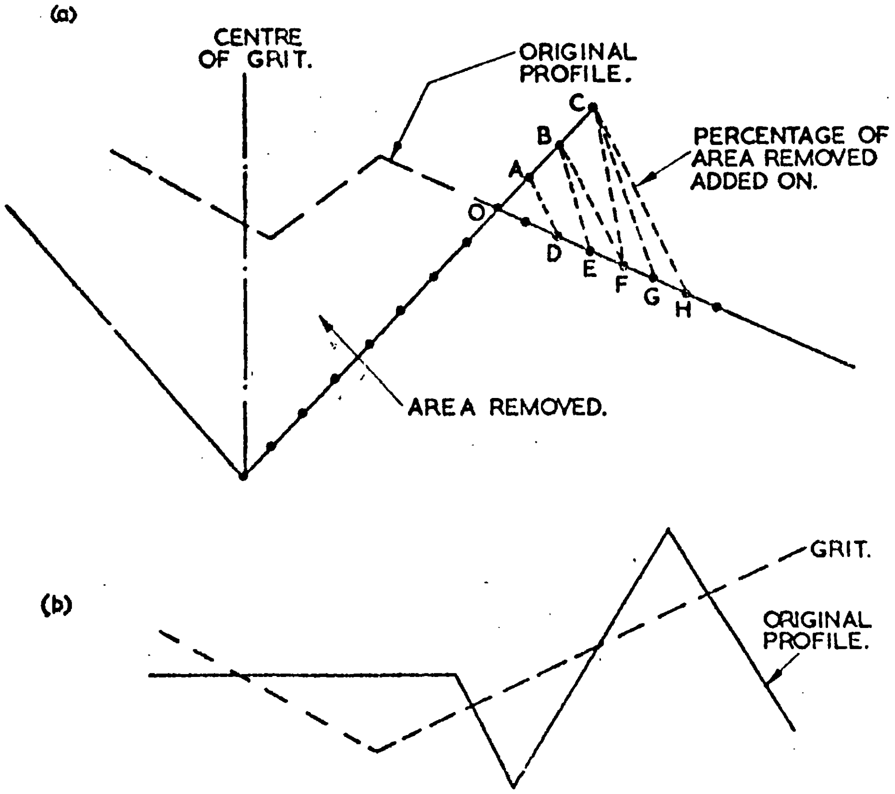
Consider, now, the conclusions to be drawn from this discussion of the effects of more complex combinations of events in the generation of a surface profile and its correlation function. Two features can be discerned. First, the replacement of events of fixed shape by events of random shape causes the correlation function to be modified from that of a single indentation; c.f. figure 7-7(a) and figure 7-7(b). Second, the introduction of random interaction also modifies the correlation function. Together these features appear to cause the correlation function to approach an exponential shape.

The theoretical analysis given above has been concerned entirely with rectangular grit shapes. The same kind of behaviour is to be expected for the more practical grit shapes. Initially the autocorrelation is that of the single event and finally, in the limit, when the degree of randomness that has been introduced is extreme then the autocorrelation function is an impulse at the origin, which corresponds to a white noise power spectrum. Under these circumstances the randomness would be such that two ordinates, no matter how close, would be uncorrelated. Random elements introduced in between these two extremes would tend, as the above analysis and the figures suggest, to make the autocorrelation more exponential in shape. It would seem, therefore, that in the same

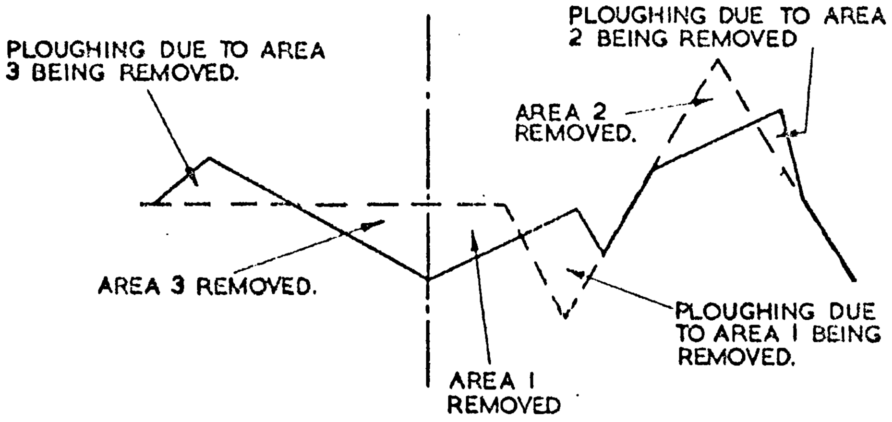
FIG 7- 13

PLOUGHING.

WHEN A CUT IS MADE THE QUANTITY OF MATERIAL REMOVED BY EACH SIDE OF THE GRIT IS CALCULATED AND THEN THE SPECIFIED PERCENTAGE OF THIS MATERIAL IS ADDED TO THE SIDES OF THE GROOVE BY THE FOLLOWING METHOD SHOWN BY THE DIAGRAM:-



HOWEVER, THE PROGRAM CAN TAKE CARE OF THIS AND THE RESULTANT PROFILE IS SHOWN BELOW.


PLOUGHING DUE TO AREA 3 BEING REMOVED.
PLOUGHING DUE TO AREA 2 BEING REMOVED
AREA 2 REMOVED.
AREA 3 REMOVED.
PLOUGHING DUE TO AREA 1 BEING REMOVED.
AREA 1 REMOVED

way that the Gaussian distribution tends to be the natural profile height distribution produced by a highly random manufacturing process so the exponential shape tends to be the natural autocorrelation function; the more the random element then the nearer it is likely to be to the exponential shape.

In order to consider these conclusions under more practical conditions further simulation experiments were carried out. These will be discussed next in Section 7-4.

7.4 Further simulation experiments

From what has been said it is clear that the shape of the autocorrelation function depends on at least two important features, the first being the shape of the unit event and the second is the nature and extent of the interaction between events when considerable overlapping takes place. There is however another feature which needs to be investigated in order to further understand the practical correlation function. This is the amount of metal movement or ploughing rather than metal removal that takes place when a grit impinges on the surface. So far in all that has been said only metal removal has been assumed. This represents a limitation of the conclusions that can be reached in terms of the practical situation. In practice, both removal and ploughing usually take place at the same time for most manufacturing processes. For example, Buttery(1968) has shown that, in grinding heat treated steels, the amount of ploughing is closely related to the hardness of the material. Figure 7-12 taken from Buttery's work illustrates

FIG.7 – 14

GENERATION OF RANDOM PROFILE FROM TRIANGULAR GRIT 25 % PLOUGHING 2500 OPERATIONS TRIANGULAR GRIT HEIGHT DISTRIBUTIONS

(a) SIMULATED SURFACE

(b) VIRGIN SURFACE

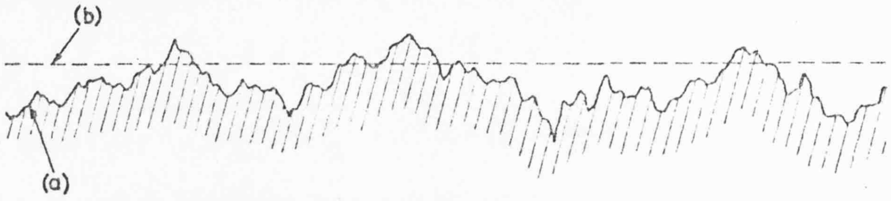
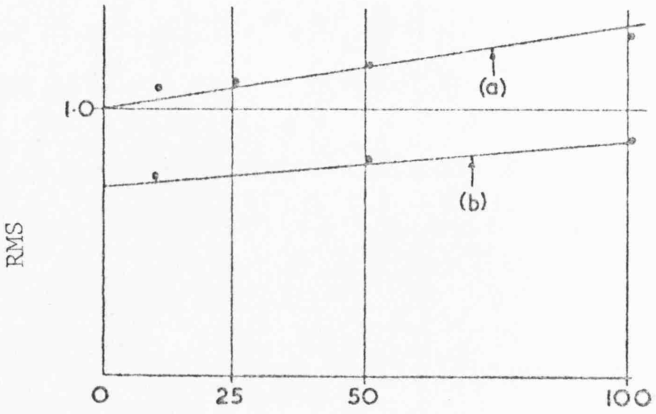


FIG. 7 - 15

RMS VARIATION WITH PERCENTAGE PLOUGHING FOR 1500 OPERATIONS. VARIABLE TRIANGULAR GRIT

(a) TRIANGULAR DISTRIBUTION

(b) SAWTOOTH DISTRIBUTION



this point. In order to take this into account the computer program was modified to perform a ploughing operation as well as a removal operation.

The program has been written in such a way that it can cater for not only the elementary situation shown in figure 7-17 but virtually any order of complexity such as are shown in figures 7-13(a) and (b). The effect of such ploughing has the effect of very quickly building up a very convincing picture of a random surface (figure 7-14).

Using the program in this form not only allows the investigation of amplitude distributions and autocorrelation functions under conditions more representative of practice but it enables some simulation experiments to be carried out on such diverse subjects as the way in which RMS roughness of the surface changes with hardness for a given number of operations. (figure 7-15). On this graph is also shown the effect of the different distributions of grit heights. As would be expected the triangular distribution of grit heights tends notonly to produce a rougher surface but also for the same number of operations it removes more overall material so one could say that as far as a finishing process is concerned the sawtooth distribution for the grits would probably be best.

As noted in the simulation experiments described earlier the amplitude distributions are very close to Gaussian, indeed with these experiments the trend is even more marked, irrespective of the shape of amplitude distribution of the grits. (figures 7-16 (a), (b) and (c)).

FIG.7-16

AMPLITUDE DISTRIBUTION OF SURFACE SHOWING EFFECT OF PLOUGHING.

1500 OPERATIONS

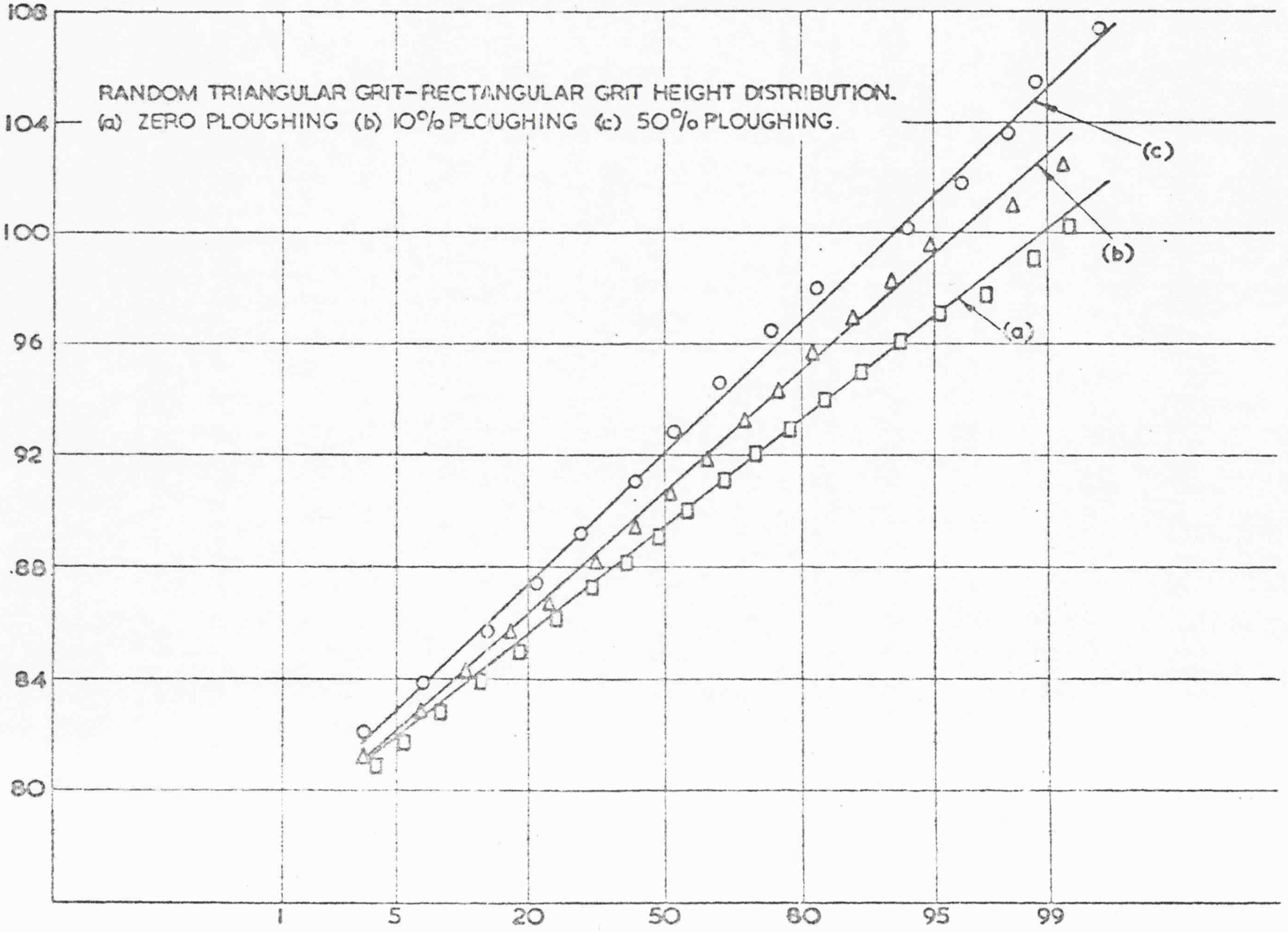


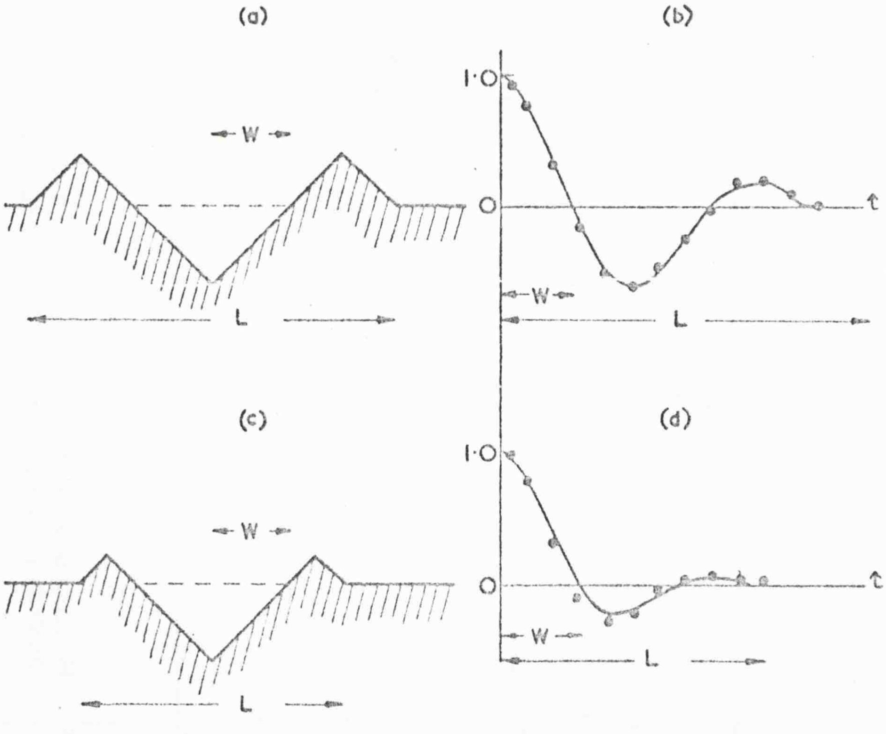
FIG. 7-17

AUTOCORRELATION FUNCTIONS GRIT IMPRESSIONS WITH NO-INTERACTION.

(a) GRIT IMPRESSION 90% PLOUGHING. (b) AUTOCORRELATION FUNCTION.

(c) GRIT IMPRESSION 22% PLOUGHING. (d) AUTOCORRELATION FUNCTION.

DOTTED POINTS ARE PRACTICAL. FULL CURVE THEORETICAL.



Figures 7-17 (a) and (b) show the simple case of ploughing where no interaction between the unit events has taken place and the unit event is a fixed width triangular shape. The examples shown are of 90% ploughing corresponding to a soft metal where very little metal is removed and 25% ploughing which is typical of results obtained with harder steels.

Immediately it can be seen that the effect of pile up on the "unit event" has caused a periodic component in the autocorrelation function. Three points of interest may be noted, (a) the extent of the correlation function length is close to that of the whole grit impression including pile up (L); (b) the distance to the first zero crossing (W) corresponds to the half width of the grit impression without pile up; (c) the magnitude of the periodic component is dependent upon the percentage ploughing.

Figures 7-18(a.1) and (b.1) show the autocorrelation functions for triangular grit impressions having random width for both 100% and 25% ploughing. In figures 7-18 (a.2), (a.3), (b.2) and (b.3) it can be seen how the shape of the correlation function of the generated profile changes as the interactions occur.

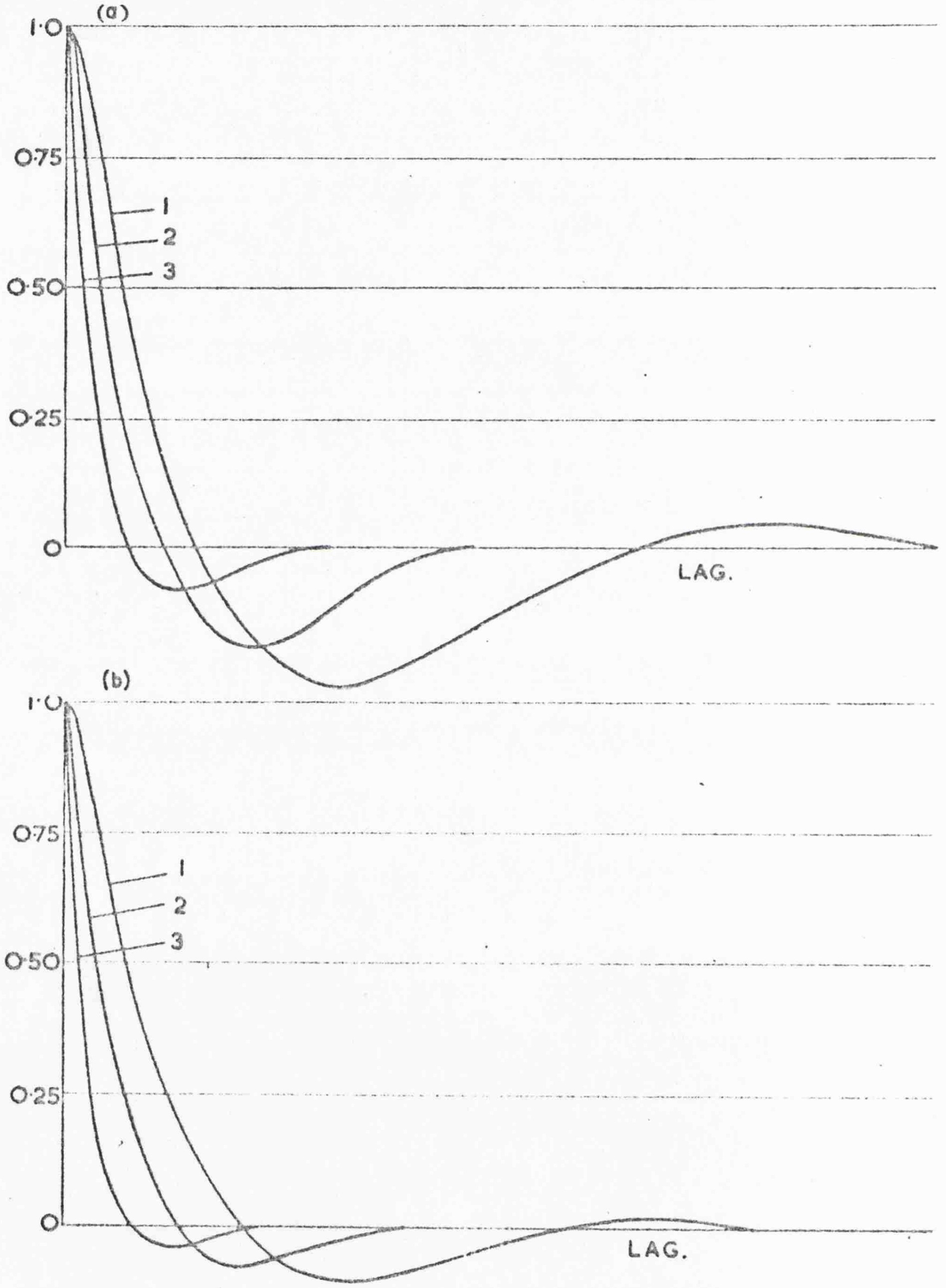
Thus it would appear that the detail of the shape of the correlation function may give detail about the mechanism of the machining and the randomness of the grit whereas measurements of size can give details about density of grits and amount of interaction.

FIG 7-18

EFFECT OF NUMBER OF OPERATIONS ON AUTOCORRELATION FUNCTION OF PROFILE WITH PLOUGHING TAKING PLACE, TRIANGULAR GRIT DISTRIBUTION.

(a) 100% PLOUGHING. (1) RANDOM TRIANGULAR GRIT 100% PLOUGHING. NO INTERACTION.
(2) INTERACTION AFTER 500 OPERATIONS.
(3) INTERACTION AFTER 7500 OPERATIONS.

(b) 25% PLOUGHING. (1) RANDOM TRIANGULAR GRIT. NO INTERACTION.
(2) INTERACTION AFTER 500 OPERATIONS.
(3) INTERACTION AFTER 7500 OPERATIONS.



As pointed out in the previous section it can be seen how the shape of the autocorrelation function tends to degenerate into the exponential shape as the random element in the shape and position of the unit events increases. Figures 7-18 show the transition from the form approximating to exponential cosine to more nearly exponential as the density of hits increases.

One thing that has emerged from this further simulation is that the autocorrelation functions for random processes which involve much metal movement and little removal such as bead blasting would be expected to show a more oscillatory shape than for hard metal grinding for instance assuming no feed marks are dominant. There is evidence (Greenwood, private communication) that this is so. This work also gives an indication that the reason why the exponential cosine type of shape occurs so often in practical cases is because of the pile-up behaviour of the unit event. Finally it suggests that autocorrelation functions regarded as having exponential cosine shape are, strictly speaking, not of this type; they have a shape worked out from the unit event mechanism (see figures 7-17 and 7-18).

7.5 Discussion and conclusions

This chapter has dealt with the way in which machining processes of a random character develop the profile geometry.

It has been seen that starting from a single scratch situation where the geometry of the profile is highly correlated with that of the machine "grit" it is possible to develop a picture of the profile in the very complex state where the random interactions of grit impressions tend to dominate the profile in such a way as to produce a more or less naturally occurring Gaussian height distribution and exponential type autocorrelation functions. As the degree of randomness (entropy) decreases then this state of affairs is less true and the shape of the individual indentations emerges. In the cases where pile-up occurs then this results in the well-known second order type of autocorrelation function resembling the exponential cosine. The horizontal scale of size of the autocorrelation function has been shown to be related also to the grit impression size and density of impressions.

Apart from the many interesting side issues arising from this investigation the main conclusions are as follows;

(a) For a random machining process the Central Limit Theorem appears to hold. Over a wide range of conditions the simulation always gave Gaussian height distributions for the ordinates. It seems to hold for cases other than those that could be easily proved statistically.

(b) The distribution of grit heights does not make a significant difference to shape of either the ordinate height distribution or the autocorrelation function but, insofar as for a given number of operations it represents different numbers of hits, then it can cause differences in the scale of size of both.

(c) The shape of the autocorrelation function depends on a combination of three things:

(i) the shape of the tip of the grit and the randomness of this shape,

(ii) mechanism of indentation whether ploughing or cutting

(iii) degree of interaction of grit indentations.

(d) For random shape and high interaction the autocorrelation function tends to become exponential.

(e) Because the correlation length is related often to the unit event, and because the shape of autocorrelation function throws light on the mechanism by which the surface has been produced, then in the much wider field of surface typology the use of the autocorrelation function as proposed by Peklenik is meaningful.

(f) One of the powers of the autocorrelation function that has emerged in this chapter, and in the friction measurements described in Chapter 6, is that it is very suitable for identifying the unit event behaviour in a clear way that is very difficult using power spectral analysis.

Summarising; the work in this chapter is justified on the basis of the considerable insight that it has given in revealing the geometrical features of the profile in terms of the details of the unit events which produce them and thereby highlights the features needed in a surface typology.

The further justification is that the work strengthens the basis upon which the theory of Chapter 4, the starting point of this work, is based.

8. GENERAL CONCLUSIONS

Developments in engineering involve progress at many levels ranging from fundamental knowledge to engineering practice and design. This thesis is concerned, primarily with the development of fundamental knowledge of surface topography and typology. Here we try to set our findings against this wider background.

The major theme of this work has been to find relationships between the functional behaviour of surfaces and their topography. In particular, we have examined the relationships between the behaviour of surfaces in contact and the features of their surface topography as revealed by digital methods. The use of digital techniques for the acquisition and analysis of data is a fairly new development in the study of surface topography but is one which shows such promise that its use seems certain to increase. In this thesis we have chosen to explore the use of these techniques in a number of different applications rather than concentrate on a more limited field.

Chapter 4 considers, in a fairly fundamental way, the problem of random surfaces and techniques used to define characteristics of significance in their contact. One immediate result that has emerged from this work is that it provides strong theoretical support for the proposals for the simple classification of surfaces by two parameters, one associated with the height of the profile waveform and the other associated with the wavelengths on the surface.

[illegible] theoretical analysis provided in this thesis. However, the justification for this any development of a simple two parameter classification must ultimately come from its acceptance and use in real engineering situations. At the same time there may well be a need for a further development of the fundamental theory to include models other than the simple Gaussian height distribution and exponential autocorrelation function used here.

For the study of random surfaces and their behaviour,the thesis shows that there exists a large body of knowledge, largely drawn from work in other branches of science and engineering which is of enormous value when applied to our chosen field. For example, in addition to the work in Chapter 4, Chapter 5 shows that the representation of a random body by means of a Markov chain can be a useful,concept providing a tool for the simplified analysis of complex subjects. The two examples discussed, one for a small body such as a stylus, and the other for a large circular body, moving across a random surface, are both practical engineering problems of some importance in surface topography. Clearly the extension of these concepts to take into account three dimensions and the effects of load is likely to be a fruitful step.

Although the experimental work reported in Chapter 6 is limited to a fairly small number of experiments, it allows some important conclusions to be drawn. First, it is clear that full, and careful, instrumentation is necessary to obtain complete information. (Our experiments are the first in which friction, ride, and changes in surface topography using relocation profilometry have all been employed). Second, once this sophisticated experimentation has been developed, it becomes clear that the stochastic approach to surface topography can be extended to include the tribological behaviour of these same surfaces. Further work in this field is clearly required and some directions in which it might develop have been suggested in Chapter 6.

If the general approach adopted in this thesis is to have overall significance it is important to be able to relate the representation of the surface profile as a random signal to some recognisable features of the manufacturing process. This problem has been tackled in Chapter 7 in a fundamental manner. It has been shown, very clearly, that the autocorrelation function of the generated profile arises from the autocorrelation function of the individual event as it is modified by other features. These features include the random nature of the individual events themselves and their interaction with the developing surface profile as the generation proceeds. These ideas seem to be of considerable importance in the development of the fundamentals of surface typology; in particular they provide a background of theory against which the classification of surfaces in terms of their autocorrelation functions (Peklenik 1967) can be judged.

Perhaps the outstanding conclusion which can be drawn from the work described in this thesis is that it highlights the usefulness and power of stochastic theory when applied to the analysis of surface topography and its relationship to the behaviour of surfaces in contact. In this respect the value of the work lies less in the originality of the stochastic techniques employed and more in their selection and application to a new field of study. Another important feature brought out in this work is that although care has to be exercised in interpreting the results obtained from the simple digital analysis of surfaces, useful results can be obtained. This thesis has tried to tackle these problems on a broad front so as to include not only the measurement and performance of random surfaces but also their generation. Therefore, some parts of the work necessarily remain to be more fully developed but it is hoped that a firm foundation has been laid for this further advance.

APPENDICES

Annals of the C.I.R.P. Vol. XVIII pp. 00–00. Printed in Great Britain 1970 MS 93 "Uncorrected Preprint"
Return to Professor Koenigsberger at General Assembly

Typology of Manufactured Surfaces

D. J. WHITEHOUSE

Research Laboratories, Rank Precision Industries, Ltd. Metrology Division, Leicester

SUMMARY. The demands of modern engineering technology are placing an ever increasing burden on the surface and surface layers of components. This problem necessitates not only a better understanding of the nature of the surfaces, but, additionally, a closer study into the methods of characterising them.

Based upon the examination of some of the more important functions to which the surface and surface layers are subjected, this paper examines the properties, both material and geometrical, which need to be typified. Additionally, this paper discusses how these properties are generated or changed by various manufacturing processes. Finally, both existing and proposed new methods of typifying surface geometry are reviewed.

ZUSAMMENFASSUNG. Die Werkstückoberfläche und deren einzelne Schichten werden durch die Anforderungen der modernen Technologie immer größeren Belastungen ausgesetzt. Die Lösung dieses Problems verlangt nicht nur ein besseres Verständnis der Oberflächenbeschaffenheit, sondern auch eine eingehende Prüfung der Methoden, die zur Kennzeichnung der Oberfläche verwendet werden.

Ausgehend von der Untersuchung einiger wichtiger Kriterien zur Beschreibung der Oberfläche und deren Schichten, befaßt sich diese Arbeit mit der Prüfung von zusätzlichen Eigenschaften, die bei der Beurteilung der Oberfläche eine Abhängigkeit sowohl des Werkstoffes als auch der Geometrie berücksichtigen. Außerdem werden Möglichkeiten angegeben, wie man diese Eigenschaften verdeutlicht oder durch Anwendung verschiedener Herstellungsverfahren ändert.

Die Arbeit schließt mit einer Zusammenstellung der bekannten und der noch im Entwurf befindlichen Kennzeichnungsmethoden der Oberflächenstruktur.

SURFACES are becoming more and more important. The requirements of modern technology are placing an ever-increasing burden on the surface and surface layers of components. This calls for a greater understanding of the nature of surfaces, of their measurement and classification, of their features, and of their control in manufacture. The requested object of this paper was to review the concept of typology, but before this is possible an understanding of what typology could comprise and of its possible application is needed. This involves a look at the nature and properties of the surfaces themselves, both physically and topographically, and an examination of what could usefully be typified in the light of their functional behaviour. This in turn involves a look at some of the functions that surfaces have to perform.

1. SURFACE FUNCTION AND PROPERTIES

The functional significance of surfaces is determined by a number of features and can be considered (a) where contact occurs with another body, either static or dynamic and (b) where no contact occurs.

1.1 *Surface contact*

Before the behaviour of surfaces under this heading such as friction and wear can be considered, a consideration of the mechanism of solid contact is essential.

The nature of the contact, either dry or through lubricant films, together with the physical properties of the materials, will determine, to a large extent, the performance of the surface; for example its susceptibility to damage in sliding contact or its ability to run-in.

Theories of surface contact are derived mainly from the equations for a single contact region usually represented by the contact between a smooth sphere and a flat surface; at light loads the deformation is entirely elastic and at heavy loads the area of contact is determined by plastic flow and is governed by the hardness of the softer of the two contacting materials. Theories of surface contact are concerned with the behaviour of individual contacts and with the subdivision of the total real area of contact into multiple contacts which occurs when rough surfaces are used.

Much of the earlier work on surface contact (Holm[1], Bowden and Tabor[2], Merchant[3]) was aimed at producing a rational explanation of Amontons' Laws of Friction which were that the frictional force was (a) proportional to the load and (b) independent of the area of contact. The first important point realised was that the real area of contact is much smaller than the apparent area calculated from the dimensions of the parts. This, together with the formulation of the adhesion theory of friction enabled the laws to be explained, because assuming plastic flow of the asperities under load L the area of contact A would be equal to L/P where P is the hardness of the softer material. Hence the frictional force would be proportional to real area of contact which in turn is proportional to the load. This theory did not attach much functional importance to the surface finish because it was assumed to be crushed very quickly under working conditions, but Archard[4] pointed out that although plastic flow could be expected to occur on the first few passes of two contacting parts in relative motion it would not continue indefinitely, some equilibrium state would occur when the asperities could support the load elastically. He then went on to show that Amontons' Laws could be explained using elastic deformation theory providing the average contact size remained constant with load. This was a direct result of having an increase in the number of contacts with load, a point which required more than one scale of size of asperity on the surface.

In order to decide which of these two deformation modes, plastic or elastic, occurs in practice it is necessary to consider the nature of the surface geometry.

The geometrical features of importance in the theory of contact are the ratio of real to apparent area, the number of contacts in a given area and the shape and distribution of the surface asperities[5]. Many investigators have tried to measure these geometrical features using a variety of techniques ranging from optical[67], thermal[8], electrical[9] to radio-active tracer methods[10]. The difficulty in measurement is because the regions of interest are more or less completely enclosed and inaccessible. However, estimates have been made for instance of the average size of the contact region and have been put between 10^{-1} mm and 10^{-3} mm, the smaller contacts tending to aggregate under tangential forces to yield the larger values. Theoretical assessment has received impetus since statistical techniques became more widely used [10–17].

One way of getting over the inaccessibility of the contact region has been devised by Williamson[124]. Using data logging facilities and accurate relocation profilometry[125] he plots contours of the surface using a computer and investigates the properties of the intersurface gap. He shows that the contact rarely, if ever, occurs at the tops of the asperities and usually occurs on the shoulders.

Another use of digital techniques has been to measure the curvature of the peaks and their distribution of heights as revealed by a stylus-type surface measuring instrument. These measurements have shown that the asperities can take on a wide variety of curvature and slope values.

One of the big problems at present being tackled is the mode of deformation of all these widely different asperities; do all the peaks deform elastically, or do they just crush plastically, or is there a complex mixture of the two modes and if so what then is the proportion of the plastic mode to the elastic mode?

In an effort to resolve this problem some workers have looked for a criterion for elastic deformation of peaks. Blok[18] showed that for sinusoidal ridges of spacing l and height $h_{\max}$ complete elastic crushing can occur if

$$\frac{h_{\max}}{l} < \frac{2}{\pi}\frac{H}{E},$$

where H is the hardness and E' is the composite elastic modulus of the two surfaces. This was modified by Halliday[19] who measured many practical slopes (θ) on surfaces and came to the conclusion that if

$$\theta < K_2 \frac{H(1-\nu^2)}{E}$$

where ν is Poisson's ratio, then elastic conditions at the peaks prevailed, K_2 being a constant dependent on the shape of the asperities taking values 0·8 to 1·7. Subsequently, Greenwood and Williamson[11] have introduced a far less severe criterion—the plasticity index ϕ given by

$$\phi = \frac{E'}{H}\sqrt{\left(\frac{\sigma}{\beta}\right)}$$

where σ is the RMS value of the surface and β is the mean radius of curvature of the peaks. For $\phi < 0{\cdot}6$ deformations will be elastic up to quite high loads whereas for $\phi > 1$ deformations will be plastic almost at a touch. Unfortunately because of the lack of definition of what is meant by a peak, this index loses impact—in effect it just gives the scale of size of asperity on any surface at which the plasticity occurs, but even this is only an average value because for one scale of size there can be a distribution of possible curvatures.

This latest point illustrates the fact that great care must be exercised in reaching conclusions based on the results of analysis of surface data, and in particular the results obtained digitally. The values of curvature for instance depend critically on the sampling rate used. Also of importance when comparing results of different workers is the need for clear definition of terms—for instance what is a peak? One final point to be taken into consideration is the resolution, both spatial and vertical, of the measuring instrument. Whitehouse and Archard[20] have considered these problems, both digital and instrumental, and have shown that unless these problems are faced the results can be practically meaningless. For instance many workers talk about the density of peaks on a surface but in fact peaks exist on manufactured surfaces of smaller and smaller size down to atomic levels. Results should be qualified with definitions of peaks, etc., and also some reference to the scale of size.

Recent work[22, 23] has indicated that, even when the proportion of plastic flow to elastic deformation has been found, the contact situation is still complex; the interaction of peaks deforming plastically can give rise to effects previously attributed to work hardening[24] resulting, for instance, in asperity persistence under crushing loads.

Turning to a more practical and extremely important point Kragelskii and Demkin[21] appear to be the first to take the waviness of the surface into account when considering the effect of the real area of contact with load. They emphasise that the presence of waviness reduces the real area of contact, the load having to flatten down the waviness elastically before the surface proper is fully contacted.

Many surface functions are determined by the degree of contact between two solid surfaces; electrical and thermal conduction, seals and interference fits come into this category. In the case of electrical conductivity the contacts provide current paths between the two bulk materials, the resistance to flow being made up of two components due to (a) the constriction of the current because of the small dimension of each of the contacts, and (b) the usually present oxide film at the contact[1]. Although the situation is similar in thermal conductivity heat transfer can occur across the gap and the effect of the oxide film is smaller[25]. Yip and Venart[26] correlate both electrical and thermal effects with surface roughness taking the very important factor of waviness into account.

In sealing both the magnitude of the gap between the two surfaces and the lay of the surface texture [27] are important. Mitchell and Rowe[28] use the ratio d/σ as a criterion of sealing for a surface where d is the average peak-to-valley distance and σ is the r.m.s. value of the surface. This ratio could be useful in predicting the sealing properties of surfaces made by single point cutting but is not so useful for surfaces that have been manufactured by random multiple point methods like grinding because in these cases, the surfaces being random, the ratio tends to be constant. Other related functions to sealing are static friction[3] press fits[29] and tolerances[30].

1.1.1 *Friction and wear*. These widely studied subjects must be touched on because they involve both the material and topographic aspects of the surface and hence must influence concepts of typology.

Considering first the friction properties. It is now generally agreed that the adhesion theory[1–3] provides the most satisfactory explanation of the observed characteristics.

The theory assumes that strong adhesive forces exist between the contacting asperities which cause cold welds at the junctions of clean asperities between metals. The presence of these adhesive forces[31] that probably vary with distance as an inverse quartic law[32] has caused Rabinowicz[33] to try to correlate adhesion with surface-free energy (which is closely related to the surface tension). If the surfaces are now moved relative to each other shearing of these junctions can occur. The adhesive theory of friction requires that two components exist, a shear component for the shearing of the junctions and a ploughing term which is the force required to push the asperities of the harder material through the softer one. Asperity interlocking is usually only a minute fraction of the frictional force, so that the coefficient of friction $\mu = S/P$ when the ploughing is negligible, P being the flow pressure and S the junction shear strength.

Factors that influence the friction between surfaces can be mechanical like the increase of friction due to the presence of residual stresses[34], crystallographic structure[35, 36], or more generally physical and chemical properties like the presence of surface films. In general these films comprise an oxide or sulphide layer formed on the metal surface[37]. Next to this there will usually be an adsorbed layer composed mainly of water, gases and grease or oil films. All these can have a considerable effect on the friction.

Temperature can affect the properties of sliding surfaces considerably. In the case of friction this is due mainly to the increase in the rate of formation of oxide films with temperature[33] which tends to reduce metal interaction[38], the mechanical shear strength and hardness terms being affected roughly by the same amount—that is until one of the melting points is being approached. Obviously also the thermal conductivities and specific heats of the materials help to determine the contact temperature reached in any situation.

From the point of view of surface geometry a number of different parameters have been put forward as having most influence on the friction. Myers[39] suggests that the slope of the flanks is the most relevant, whereas in the case of the friction of metal on polymers, Trott[40] shows that the curvature of asperities is important—presumably because of the interlocking effect. The slopes of the asperities should be only marginally important in practice because of the dominance of the adhesion term, although measurement of the variation in frictional force can in certain circumstances be a measure of the first differential of the surface geometry. Kragelskii[31] uses the bearing area curve to work out friction from a profile graph.

A recent comprehensive review of wear has been given by Archard[41]. Wear is defined as the removal of material from solid surfaces as a result of mechanical action (although some engineers would also include any displacement of metal that produces a dimensional change).

Whereas in the past wear usually continued through the roughness layer because of the imperfect manufacture of the surface, recent improvements have been such that the wear throughout the life of the component is often confined to the outermost layers which are therefore becoming increasingly important.

The major difference between wear and friction is that all points of contact make a contribution to friction, but a contribution to wear is made only by those contacts which result in damage or metal removal when the junction has sheared. This number is only a small fraction of the number of contacts, perhaps one millionth. Whereas friction is dominated by the behaviour of contacts which involve virtually undamaged surface layers, wear is concerned only with those few in which the surface layers are broken.

The simplest classification of wear originated with the friction classification of Bowden and Tabor[1]. This classification was into "severe" and "mild" regimes. In severe wear the contact is largely metallic. The surface damage is great and large particles up to hundreds of micrometres are formed whereas in mild wear the surface is relatively undamaged and protected by oxide films (Hirst and Lancaster[42]), very much smaller particles being produced. According to Archard[41] this classification is primarily a distinction in the scale of size, severe wear involving contact sizes of 10^{-1} mm whereas mild wear might be several orders less.

The individual mechanisms in the wear process have been classified in a reasonably straightforward way by Burwell[43] as follows:

Adhesive wear is primarily in terms of metallic welding and can probably be taken to represent the first stage of the severe wear regime. It involves the transfer of metal from one surface to another[44] and then the release of some of this transferred material as wear debris due to either chemical action which reduces adhesion or to residual stresses in the transferred material, Kerridge and Lancaster[45]. Rabinowicz[33] maintains that a transferred particle comes off only if its stored elastic energy is greater than its surface energy.

Abrasive wear is due to the cutting or ploughing of a hard surface or particle into a softer surface, called "two-body" if only two surfaces are involved and "three-body" if loose debris particles are also involved. Kruschov and Babichev[46] consider the wear resistance of metals by taking into account their relative hardness. They find that the resistance to wear is large if the surfaces are harder, although work hardening is not important.

Corrosive wear requires the presence of both corrosive agents and rubbing, in particular oxides and hydroxides of metals are easily formed in corrosive environments and since they are usually loosely adherent even mild rubbing can remove them.

Fatigue wear or pitting occurs primarily when surfaces undergo cyclic stress patterns such as might occur in rolling or reciprocal sliding conditions.

Although pitting usually occurs under conditions of partial or full lubrication it is convenient to refer to it in the section on wear. Dawson[47] has shown that the criterion of pitting is highly correlated to the ratio of the total height of the surface roughness to oil film thickness. If this ratio is less than unity, pitting is not likely to occur. Scuffing is likely to occur in some situations where pitting is possible, particularly when the slide-to-roll ratio is high. This is a condition associated with the high temperatures generated during contact.

Another common form of wear is fretting[48], where contacting surfaces undergo small tangential displacements of small amplitudes.

In general wear, as friction, is influenced by speed, temperature and load. Under low speeds, low temperatures and low loads mild wear occurs whereas at large values, severe wear occurs.

1.2 *Intermittent solid contact*

Consider the influence of surfaces on lubrication. Depending on the lubrication regime, the surface effects are of different relative importance. For instance, in hydrodynamic lubrication the surface contact and hence possibility of wear, apart from the intermittent contact due to debris during running, occurs only at start up and shut down. The debris particles, unless filtered out, cause further wear and a progressive deterioration of the oil film thickness. Tallian *et al.*[50] and Furey [51] have attempted to correlate the amount of wear with the degree of metal contact as revealed by their experimental procedures with some success. Christenson[52] considers contact in the regimes between boundary and hydrodynamic lubrication and investigates not only the number of metal contacts but their average duration and the rate of variation, parameters very relevant to topographic typology. In the hydrodynamic and elasto-hydrodynamic lubrication regimes it is the bulk properties that usually decide the conditions, whereas in boundary lubrication it is the surface layers[31] which either take the form of a molecular layer of lubricant on the surfaces or of a solid film such as an oxide or sulphide. Rabinowicz[33] maintains that a property of surfaces that helps the boundary lubrication properties is the surface energy which when high facilitates the adherence of lubricant molecules. The surface should also have energetic compatibility with the lubricant (Imai and Rabinowicz[53]), that is, it should be capable of being wetted by the lubricant.

1.3 *Fatigue*

In several cases the surface plays an important part and yet does not directly make contact with other surfaces, for instance, in fatigue. The fatigue cracks often start at or near to the surface layer in regions of high tensile stress. This sort of failure may therefore be caused more by the stresses produced in manufacturing the surface than by the surface geometry itself. It is the release of these residual tensile stresses that causes fracture of the surface layer and hence gives a basis for the propagation of fatigue cracks.

Fatigue can be caused by a number of other surface phenomena such as deep surface scratches, crystalline dislocations, etc. Another effect of internal stress can be to distort a part from its true form. From this it is clear that stresses in the surface layer can influence how a part behaves. These stresses according to Van Hasselt[54] can be produced because of plastic deformation from non-uniform thermal expansion, volume changes from chemical reactions, precipitation or phase changes and mechanical working. In practice in order to get rid of unwanted tensile stresses the part is sometimes shot-peened to give a resultant compression in the surface[55, 56]. It then seems likely that residual stresses are more important than surface finish although the finish still plays an important part.

1.4 *Bonding contact*

Coming under this heading is the coatability of surfaces, which concerns the covering of the surface with a layer of material either for decoration or anti-corrosion as in painting and plating. The coating can be either chemically bonded to the surface or physically bonded as in diffusion. Here obviously the chemistry of the outermost layers is of prime importance in determining the degree of bonding, also the adsorption of grease and oil, Adam[37]. In addition the surface finish can affect the rigidity and appearance of the coating, and it can determine the volume of paint necessary to paint the object[57]. Surface energy is likely to be important as are the electrochemical properties of the surface layers. Stresses are also important because they can cause cracks which expose metal to the atmosphere causing adverse chemical or galvanic action.

1.5 *Other forms of interaction*

Into this category may be put interaction with waves, gases, fluids, etc. The scattering of electromagnetic waves, in particular light waves, by surfaces is an important functional consideration. Much work has been carried out in this field not only from a visual point of view, but also in the investigation of the nature of the surface of the moon and sea[58].

Theories of backscatter have been advanced for both scalar analysis[59], and vector analysis[60]. Fraiture[60] shows that autocorrelation functions and r.m.s. values of the roughness can be found, theoretically at least, by measurement of backscatter. Other surface information can also be extracted such as the average slope of the asperities. Bennett and Porteus[61] have evaluated the r.m.s. roughness from specular reflectance at normal incidence, although only for fine surface finish where the slopes can be neglected. Most investigators, like Davies[62], assume perfectly conducting surfaces and uniform incident reflectance, properties which together with the surface geometry are of the utmost importance in questions of appearance like sheen gloss[63]. Dielectrics do not have reflectances as high as metals but information can still be obtained from them.

Richmond and Steward[64] have shown that the spectral emittance of metals can be increased by two or three times by roughening the surface. This effectively increases the absorptance and hence, by Kirchoff's Law the emittance. For dielectrics this is not so marked[65]. The reflection of coherent light is affected by the geometry of the surface; for a given numerical aperture the speckle intensity variation depends on the bandwidth of the roughness[66].

Functional effects involving the contact of gases or fluids are too numerous to go into detail, but take as an example of gases the simple effect of the atmosphere which can cause corrosion of the parts —especially if the surface has high residual stresses which can cause bare metal exposure, and another example is the presence of chlorine which can affect fatigue. The passage of fluids or gases can be influenced by the geometry of the skin, too rough a surface producing local turbulance and frictional heating, e.g. on aircraft wings.

Summarising, it is obvious that in many practical situations the surface and surface layers of a component can determine to a large extent how well it behaves in working conditions. The functions here considered, although by no means a complete list may provide a sufficient sample to indicate which may be the important properties and hence what needs to be typified.

How a surface behaves in practice when subjected to various loads, speeds of relative motion, temperatures and environments is likely to be determined by not only the topological features such as peak distribution, average height (R_a), autocorrelation function, isotropy, waviness, etc. but also the material properties which may be either mechanical, physical or chemical, or more than likely, a mix.ure of all three. These surface properties would have to include the hardness, flow pressure, the modulus of elasticity in both shear and tension, crystalline condition and crystal orientation, specific heat, thermal and electrical conductivity, reflectance, surface energy, internal strain and stress conditions, chemical activity and adsorption, homogeneity and impurity content to mention just the obvious properties.

It could be thought that most of the material properties would be inherent in a statement of the material to be used for a particular purpose. Unfortunately this is not likely to be so because the properties on the surface can depend so much on the manufacturing process.

Hence this shows what an enormous amount of information may be involved in an effective Surface Typology. Further, as Field and Kahles[67] point out, both material and topographic conditions must be specified and secured if the proper functioning of the surface is to be assured.

2. SURFACE TOPOLOGY

Up to the present surfaces have been typified in one rather rough though inclusive way by specifying the manufacturing process in conjunction with a simple numerical index such as R_a or r.m.s., the process in effect determining the basic material properties as well as the surface geometry, the index merely ensuring some degree of control over the satisfactory application of the process.

How has this system worked out? To try to decide this the manufacturing process itself must be examined to see how good a basis for typology it really is.

2.1 *Manufacturing process*

Any form of metal removal is liable to cause differences between the surface and the bulk properties. These differences are usually caused by high temperature or temperature gradients developed during the removal process, plastic deformation and flow of metal, and chemical reactions due to the presence of impurities, coolants, etc.[67]. These differences between surface effects and bulk properties are being found increasingly important with the use of nickel, titanium and other alloys, and also because materials are being heat-treated to higher strengths. The principal alterations caused by these conditions are (1) change in hardness of the surface layer, (2) recrystallization or phase transformations in the outermost layer leaving for example hard martensitic aggregations in the ferrous alloys, (3) residual stresses, (4) embrittlement by chemisorption, and others. A brief review of how these effects are brought about in the different processes may be justified.

In cutting processes like turning, two sorts of finish can result, clean or torn depending on whether a built-up edge forms on the tool[68] and on whether or not micro-chips develop on the tool flank[69]. The build up on a tool edge depends largely on the tool edge temperature which in turn depends on cutting speed. For high cutting speeds the built-up edge disappears due to metal phase transformations, the surface layers are subject to higher temperatures and the roughness becomes independent of rake angle, depth of cut and speed[70]. On the other hand at slow speeds the surface layers are subject to tearing and more severe work hardening.

The mechanism of chip formation is of prime importance in cutting processes, determining not only the stress pattern left in the surface layer but also the heat generated. Another factor is the variation in cutting and frictional forces which can accentuate if not initiate chatter[71].

Metallurgical consequences and flow properties of the cutting action are dealt with by Turkovich and Calvo[72] and Turkovich and Micheletti[73] who show that most of the heat generation actually taking place in the flow zone is due to the creation and destruction of disolocations.

Because of the method of metal removal, the shear zone in cutting is likely to contain tensile stresses due to the flow of the hydrostatic component of metal into the chip and also the type of chip fracture. The rake angle of the tool can influence the surface layers; for instance for small positive angles the average thickness of the chip increases and the depth of the plastically deformed shear zone in the surface increases thus increasing residual tensile stress concentrations[74]. A typical depth of the shear zone is about fifty times the surface roughness.

In grinding the grits can be considered to be single point tools of random rake angles having random position relative to each other, so they do not as a rule give the same result as that of a simple cutting process. Also it is possible for many grits to pass over the same workpiece position for one wheel revolution, producing overgrinding, which introduces random thermal and mechanical stresses[75]. Another point to notice is that the process of grinding can be self-dressing in the sense that as the grinding proceeds, grits fracture and present new edges to the component. This only happens if the part speed and wheel speed are suitable.

Generally temperatures generated during grinding are higher than in single point cutting, so phase transformations and recrystallizations are more likely to occur, but on the other hand plastic deformation is lower resulting in lower internal stresses. Geometrically the texture in grinding is finer than in cutting operations and has as a rule a Gaussian ordinate height distribution, which is not necessarily the case for single point processes. One factor that can greatly influence the surface finish of the part is the presence of chatter which produces waviness[76]. Farmer *et al.*[77] found that to get best surface finish the table speed should be kept low. Thorough removal of the wear debris during the process also leads to better finish[78].

Recent work has indicated that to get a clear picture of the grinding process, forward as well as lateral flow of metal have to be considered. Also not only chips are produced but slivers resulting from the wide walls. These and other effects have shown that grinding has differences from the conventional cutting techniques[79]. Energy losses are far higher in grinding than in cutting because of higher contact temperatures.

Abrasive techniques involving an area of contact (e.g. honing and lapping) are still not fully understood. They differ from normal grinding and cutting processes, chiefly because large areas are contacted at low speeds[79]. Polishing is different again[80] because the abrasive is held in a very soft bond and is therefore relatively free. Samuels does not subscribe to the theory of Bielby[81] which postulates the formation of an amorphous layer of molten metal on the surface during polishing. Instead he favours the abrasive theory which does not preclude asperities of all scales of size being present. One of the big differences between polishing and cutting is that in polishing abrasive debris can easily become embedded in the surface and be almost impossible to remove.

Metal forming methods involve a considerable amount of plastic flow which may greatly affect the crystalline structure and topography of the surface, but are too extensive in their variations for consideration in the present paper. However, even when these methods are not the finishing process they may have to be taken into account because they can decide the material property history of the surface.

Electro-machining of metals also produces its own variety of problems, for instance in spark machining thousands of amps per square cm may pass through the surface producing both physical and chemical changes. In hardened steels, for instance, after machining there is a thin resolidified extremely hard layer, often much harder than normally possible, beneath which there may be a thin tempered zone. Also high tensile stresses down to at least twice the surface roughness depth can exist. Electrochemical machining at low current densities can cause intergranular cracks in multi-component alloys. From the topology point of view in electro-spark machining, rough surfaces are usually produced while electro-chemical machining can get fine finishes.

The preceding sections have tried to show how the manufacturing process can influence the material and geometric properties of a surface and consequently influence all those functional behaviours mentioned earlier, and thus react on the general problem of typological specification.

Although a number of processes are usually used in the stock removal and finishing operations involved in the manufacture of a component it is not only the types of process that are important in determining the surface properties, but also the *order* in which they are applied—thus giving different thermal histories.

Again, as just mentioned there can be variations of surface properties and topography within any one process. A point brought out by Kahles and Field[67] is that considerable differences can arise in the material properties of a surface resulting from gentle and abusive treatment; on the other hand Olsen[83] evolved typical good conditions for turning. This seems to show that specification by process may not always be adequate; it also shows how great a problem will be set in the evolution of a fully sufficient specification by typology.

When considering the shortcomings of specification by process, even by process specified in detail, two further points arise:

(a) Specification of the process may not guarantee against its misapplication, and it would be desirable if surface property tests could be made after manufacture as a form of inspection, although this may be a far-off objective.

(b) The production engineer wants freedom of use of production plant. He does not want to be tied down to any one process for a part because this could disrupt his scheduling of jobs in the workshop.

Consequently, the modern search is for typology both of the material and geometrical properties. Numbers indicating the requirements would be put on the drawing and the job of the production engineer would be merely to reproduce the properties corresponding to the numbers in any way he pleases. This is the theory; the solution is not easy.

2.2 *Material property typology*

To specify the material properties adequately it would be necessary (in addition to specifying the material itself) to control the mechanical properties of the material, e.g. its elastic modulus and the hardness values at the surface itself, also the degree of residual stress allowable in the surface layers. Surface energy and chemical properties might be involved although they might well be intrinsic in the specification of the bulk material itself. It might also be true to say that a fairly detailed metallurgical specification could to some extent anticipate the mechanical properties. It could be that the complete typology of the material properties will be too difficult to achieve and that the nearest approach will be something like that of Greenwood and Williamson, who aim to predict the functioning of the surface by including in their plasticity index both material and surface topography parameters.

2.3 *Topographic typology*

The typology of the surface geometry also presents problems and continues to defy complete solution. This is mainly because of the complex character of the surface geometry and also of the multiplicity of different functional uses to which the surface could be put. In an effort to provide data from which a typology could be developed, a joint OECD/CIRP research programme was launched in which typical manufactured surfaces, prepared by the Technical High School, Aachen, (Professor Opitz) were circulated to different institutions and firms through the world for evaluation and classification. In the Rank Precision Industries Laboratories, their profiles were recorded digitally on tape and basic parameters were then computed. Unfortunately not enough specimens were provided of each sort of manufacturing process to enable any sound conclusion about typology to be made from the results, which nevertheless gave useful information. More of this sort of co-operation, however, is essential if the problem is to be solved.

Discussion of typology leads naturally to statistical considerations. The study of these statistical parameters has been greatly facilitated by the use of digital techniques introduced by Reason[85].

Surfaces can be random or deterministic in character or more usually a mixture of both. For a complete specification of a general random process, high order joint probability density functions are needed[86]. However, in practice a second order joint probability density function $f(y_1 y_2;\, x_1 x_2)$ is sufficient where $y_1 y_2$ are ordinate heights and $x_1 x_2$ are spatial co-ordinates. From this the ordinate height distribution (or strictly the ordinate height probability density function) can be obtained as a marginal density function by integrating $f(y_1 y_2;\, x_1 x_2)$ with respect to y_2 or y_1. This gives information about the relative frequency of ordinates at any given height. Also, the autocorrelation function can be obtained, because it is the joint moment of $y_1(x_1)$ and $y_2(x_2)$ and gives information about the way in which the ordinates follow on from each other. The autocorrelation function is usually normalised with respect to the variance (mean square value) of the signal. A direct equivalent of the autocorrelation function is the power spectral density function which can be obtained by a straightforward Fourier transformation. Because the second order probability density function is in general not known the autocorrelation function and ordinate height distribution can be used to define the statistics of the profile. In the many practical instances where the statistical process is Gaussian or thereabouts then the normalized autocorrelation function and the r.m.s. value (or average value R_a) completely define the profile (assuming it has zero mean value)[87]. The autocorrelation function has the useful property of being able to separate the random from the periodic components of a waveform, although because of its phase destroying property it cannot tell whether or nor the periodic component has random phase elements. Standard methods of obtaining these statistical parameters have been well documented[88, 89].

The usefulness of correlation and spectral analysis techniques has been appreciated for many years in other fields including medicine, where they have been used for analysing electroencephalograms[126], for oceanography[118], and for siesmology[127], to name just a few applications.

Most of the present day parameters of surface geometry are basically estimates either of the height distribution or of the autocorrelation function or mixtures of the two. A practical requirement inflencing the choice of parameters is needed to keep instrumentation costs low and techniques relatively simple, furthermore, for a parameter to be useful in typology it must be at one and the same time both discriminatory to distinguish between one surface and another both produced by the same process, while being steady enough not to produce wildly varying values over the same surface.

Another point which tends to be neglected is that any topographic typology should be capable of taking into account not only the overall statistics of the surface but also statistically unpredictable freak events, such as the odd deep scratch which (assuming it can be located) may be important in functions like fatigue. This is one region where the use of correlation techniques is no benefit at all; similarly the prediction of extreme behaviour is not at all straightforward[93]. Such behaviour is by its very nature difficult to predict—even using statistics.

The importance of the ordinate height probability density function in surface metrology was first realised by Abbott and Firestone[94] who proposed the use of a curve showing how the ratio of metal to air changed with the height of a hypothetical flat plate lapping away the surface from the highest peak to the lowest valleys. This curve is generally referred to as the bearing area (or ratio) curve; it is in fact unity minus the ordinate height distribution function. Pesante[95] proposed topological classification according to the shape of the ordinate height density function. He proposed that the density value at a given height should be taken and, later on, that it should be augmented with a peak count. Pesante found the density function more useful than the bearing area curve because it is inherently more discriminating. Reason[96] proposed the use of the consolidated bearing area curve together with the high-spot count to classify the surface, and Ehrenreich[97] suggested that measurement of the slope of the bearing area curve could be functionally useful. The CLA (R_a) and r.m.s. values are essentially estimates of the scale of size of the ordinate height probability density function—as indeed is the maximum peak-to-valley height and ten point height. Other features of the density function can also be considered useful, especially in demonstrating wear, for instance the skew. Al-Salihi[98] in fact proposes that, in addition to the scale of size estimate, the third and fourth central moments representing skew and kurtosis should also be relevant. Taken as a whole, these parameters would provide a much more comprehensive estimate of the density.

Unfortunately, measurement of the higher order central moments of a distribution are not as a general rule reliable because of the tendency of the value to be dominated by the rare very large peak or valley. Unless some rule is used which excludes these freaks the results can be doubtful. One point relating to the shape of the bearing area curve (or the density curve to a lesser extent) is that it is not very discriminating. Over many surfaces encompassing a variety of processes, the curves hardly change and this feature, taken with the known wide possible variations in instrument values, Von Weingraber[99], make it rather likely that any significant differences could be swamped. The fundamental reason why the density function by itself is of limited value is that it contains no information about the bandwidth of the profile waveform. Many people have pointed out[100] that there is a functional need for a spacing type parameter, for instance in the sheet steel industry [101]. Hence, to stand any chance of being a suitable basis for a system of typology the ordinate height density curve, or an estimate of it, has to be complemented by the specification of another feature such as the process of manufacture or the autocorrelation function for instance.

Any typology must be a compromise. This is not surprising when it is remembered that only a few dozen bits of information from all the millions present in a typical profile waveform are going to be significant for any given function. The question is, which bits? This is where the autocorrelation function is useful because it represents a useful condensation of the information in the profile waveform without losing information about the energy in the component waves making up the waveform. Wormersley and Hopkins[102] were the first to put forward the autocorrelation function (in the form of a time series) as a useful measure of surface texture, followed by Linnik[103] and Nakamura[104]. However, it was Peklenik[105] who proposed the use of the autocorrelation functions for the specification of the typology of surfaces, and the Author has derived much benefit from his contributions. He proposed classifying the autocorrelation function into five different groups; a surface being typified by examining the shape of its autocorrelation function to decide into which group it best fitted. The surface was then typified by the number of the group.

Thus he was able to present surfaces made by different processes on a typographic scale which comprises:

Group 1—Cosine or steady.

Group 2—Exponential decay plus cosine.

Group 3—Exponential decay modulating a cosine.

Group 4—Complex combination of Groups 2 and 3.

Group 5—Exponential decay.

In this classification, Group 5, for instance (first-order random surfaces) is typical of those manufacturing processes best described by Poisson point process statistics[106] such as grinding, honing etc. whereas Group 3 (second-order random surface) together with Group 2 are more typical of single point cutting processes like shaping, turning, etc. Group 1 is purely deterministic and does not occur on practical surfaces.

Three points are valid in the discussion of these important concepts, the first being that the reliable measurement of autocorrelation functions is not ve.y easy. It is usually accepted that correlation functions are better used for the estimation of the presence or absence of features rather than of their quantative values[107]. The use of double correlation can sometimes help in this matter[108]. The second point is that a more comprehensive typology scheme could be formulated by classifying the shape of the ordinate height density function *in addition to* the autocorrelation function. Functionally a classification of the ordinate height density function need probably only take the skew into account, for example, +1, 0 and −1, depending on whether the density function is positively skewed, symmetrical, or negatively skewed. It would be necessary to define symmetry, the zero class, by imposing a certain limit say ±0·2 for the skew within which the distribution would be classified as skew zero. These three groups taken with the last four groupings of the autocorrelation due to Peklenik would provide a comprehensive classification and would be reasonably identifiable. That a classification based only on the autocorrelation function can be sometimes insufficient can be demonstrated easily. Take for example the random telegraphic signal which consists of a signal switching at random times between two fixed levels. This has got an exponential autocorrelation function and yet it looks nothing like a typical lapped surface having an exponential autocorrelation function.

Finally, there is the point that the extracting of the d.c. level and curvatures usually present and unwanted on a profile graph can cause a change in the autocorrelation function. For instance with Group 5 surfaces the removal of these misleading low frequencies can change the appearance of the autocorrelation to that similar to Group 3. However, with care, the group classification can still be of use.

As a further subdivision of each group, Peklenik introduces the correlation length and the correlation period, the former being that separation of two points on the profile that makes them just independent of each other, and is measured by the lag distance in which the autocorrelation function dies finally to a fixed value, between 10 and 50% of the original. That is, the conditional joint probability density function becomes equal to the ordinate height density function. The correlation period is the wavelength of the dominant periodicity in the autocorrelation function. Again these subdivisions are useful but sometimes they are made difficult to measure by the presence of unwanted features; for the correlation length by random fluctuations in the autocorrelogram, and for the correlation period by harmonics. Moreover, as mentioned above the extraction of the unwanted very low frequencies presents problems.

Whitehouse and Archard make two significant contributions. The first is by linking together the parameters used in the classification system of surfaces (due to Peklenik) with the parameters directly useful in the functional assessment (due to Greenwood and Williamson). The second point is that they analyse the variations in parameter values that can arise when using digital techniques.

After proposing just two values to completely represent a random waveform, the r.m.s. value or average (R_a) value the correlation length they derive expressions for the distribution of peak values, curvature at peaks and the plasticity index[11] using only these two geometric parameters. They show, for the representative case of surfaces having a Gaussian ordinate height distribution and an exponential autocorrelation function, that the peak distribution becomes more Gaussian with height and that the peak curvatures approximate to a Rayleigh distribution. Also they show how, when using digital techniques, large differences in parame er values can result depending on the sampling interval used, and from this decide on limits for this interval which, in turn, leads to a way of determining the broad structure in the surface.

Correlation length concepts have also been used by Beckman and Spizzichino[59]. In addition, purely spectral methods of classification in surface finish have been used by Ber and Braun[110] and Dunin Barkovsky[111], who mentions cases where it has been useful functionally, although these examples are not properly explained.

Turning now to other methods of typology, one important attempt was made by Myers[110] who recommended the use of the r.m.s. values of not just the profile itself but the r.m.s. values of the profile slope and second derivative, together with a directional parameter.

Other investigators have proposed the use of either one or more of the derivative parameters. Peklenik[113] considered the value of the standard deviation of the slope as a convenient estimate of the autocorrelation function, the function itself being costly to evaluate directly because of the complexity and cost of correlation equipment and the length of time necessary to produce the correlogram. However, this is becoming less of a restriction than previously because of the advent of fast Fourier transform techniques which can considerably speed up operation and accuracy[114]. The significance of the standard deviation of slope rather than the average value of slope, is that the slope variance is equal to the curvature of the autocorrelation function at the origin. Peklenik then suggested that the measurement of level-crossing probability is a useful practical way of determining the slope standard deviation, thus completing the link back to the correlation function. A variation on this level-crossing technique is to measure the average peak thickness at any level. The use of the distribution of slope has been reported by Kubo[115], Nara[116] and others. Nara[116] suggested that the R_a value and the mean slope value could be used for specifying a surface on a two-dimensional graph. Using this information he then represented the surface by an assembly of cones of half angle given by the average slope. This he found useful in specifying electrical resistance. Using this model he was able to make estimates of both the correlation length and the high spot density.

To get to a very important practical point concerned with most of these discriminating parameters like slope or curvature measurement and level-crossings measurement, they tend by their very nature to reduce the effective signal-to-noise ratio, i.e. extraneous short wavelength noise tends to get amplified relative to the required signal—a natural consequence of this being that the value obtained for the parameter can be dominated by the noise. One way out of this problem is to introduce a short wavelength filter to cut out this noise. However, it is of limited value to apply an arbitrary cut to the spectrum and it is best to try as nearly as possible to relate the position of the short wavelength filter cut-off to the spectrum of the surface in such a way that what is left will be functionally significant as far as can be ascertained, and not just the result of an instrumental convenience. Spragg and Whitehouse[117] have not only done this by referring to some of the theory developed in[109], but they have proposed a unified system of surface metrology based on a typology that includes the well accepted R_a, together with a parameter they have called the average wavelength and a skew index based on peak values. The average wavelength as they define it is based on a second moment estimate of the power spectral density function and is a significant measure for both random and deterministic signals. Thus, this parameter often reveals the value of the tool mark spacing for the large class of surfaces coming approximately under Peklenik's Group 3—(second-order random surfaces like milling, turning, etc.) even when the periodic component is not visible. Fortunately this average wavelength parameter can be obtained instrumentally without too much difficulty. Another feature of this system is that it can be applied equally to roundness and waviness, etc., which makes it suitable for an overall surface geometry typology. Finally, the unified system has the advantage of being able to indicate whether a R_a value that has been obtained is based on a suitable meter cut-off.

All the methods so far have dealt with the typology of surfaces two-dimensionally which means that the assessment is usually taken across the lay where the bandwidth is greatest and requires the least distance of traverse to get a reliable answer. Obviously if only one track is to be taken then this is sensible—as it must also be for isotropic surfaces, but there are occasions when a fuller description is necessary. This requires typology in three-dimensions and not only in two. In the field of three-dimensional topographic analysis much work has been done, in particular by Longuet-Higgins[118] in oceanography. Another field where three-dimensional analysis is important is in geography[119] and mapping.

McAdams *et al.*[120] have applied hypsometric (area-altitude) analysis to random surfaces such as abrasive surfaces. He breaks down areas of the surface into linear combinations of basis functions; each function having a simple geometrical shape. It is claimed that breaking down the surface into these forms of elemental shapes enables a more direct tie up with abrasive particle geometry to be made than can result from correlation techniques.

As in the two-dimensional case, Peklenik and Kubo have been prominent in the three-dimensional considerations also. They propose for the three-dimensional assessment of a surface two methods: (a) use of cross-correlation techniques for parallel tracings along the surface and (b) plots of correlation lengths on a polar diagram for radial tracings. In the parallel tracing method[121] information about the persistence of waveform is obtained by measuring the maximum value of the cross-correlation of separated traces, whereas in the radial method[122] which is more suitable for weakly directional surfaces, the degree of isotropy can be indicated on a polar plot of the correlation lengths.

Both papers bring to notice a very important aspect of surface metrology, namely the complete three-dimensional classification of surfaces. However, as yet, the amount of effort required to obtain this sort of information is large, and more work is required in this field before practical methods are likely to be evolved.

Another, or additional method for topographic control would be similar in effect to the use of a visual Atlas, as is used in metallurgy and was suggested for surface typology by the late Professor Bickel, the eye being used to compare pictures and profile graphs of a reference surface with the surface to be classified. This method has the advantage that the particular geometrical features of the surface that make it acceptable do not have to be known explicitly.

3. CONCLUSIONS

A few final points about this typology, both material and topographic, must be made.

(1) Everything that has been said refers to the surface as made. What is really needed is something predictive in nature giving information about the final surface which will not exixt until after the surface has been put to use. Ostvik and Christensen[123] show how quickly a surface changes with running-in. Strictly speaking any typology should be aimed at an assessment of these run-in surfaces and not of the virgin surface.

(2) Judging from the foregoing review, it is doubtful if there can be an effective typology that does not take both material and topographic aspects into account.

(3) The use of the manufacturing process as a basis of typology will only be really effective if the specification is made in some detail. Otherwise abusive manufacture could make functional control difficult.

(4) Waviness should not be neglected in geometric typology because the functional behaviour of mating surfaces depends largely on the extent of contact which, in turn, depends on the waviness.

(5) No system of typology is likely to find its way into industry unless it can be reduced to terms capable of being understood in the workshop, and leads to demonstrable commercial benefits sufficient to justify its cost of operation. It is possible that this requirement will impose fewer restrictions when the system is associated with fully adaptive control.

From what has been said it is clear that more work needs to be done before an effective typology can be formulated. This paper has tried to show the extent of the problem and indicate some of the lines along which work on typology is now being pursued.

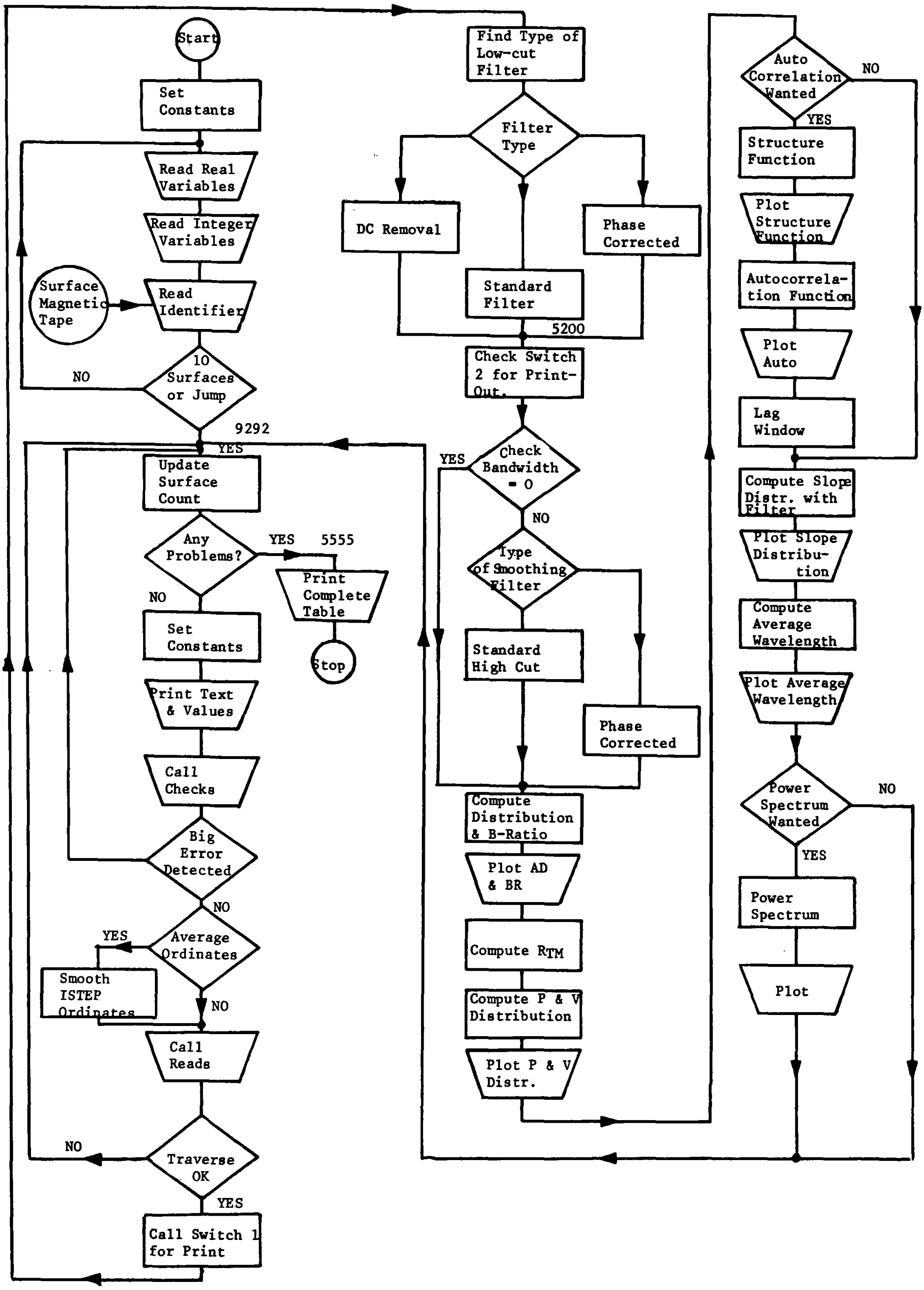
REFERENCES

1. HOLM, R., *Electric Contacts Handbook*. Springer Verlag (1958).
2. BOWDEN, F. P. and TABOR, D., *The Friction and Lubrication of Solids*, Part 1. Oxford University Press (1954).
3. MERCHANT, M. E., The mechanism of static friction. *J. Appl. Phys.* **11**, 230 (1940).
4. ARCHARD, J. F., *Proc. Roy. Soc.* **A243**, 190 (1957).
5. LING, F. F., On asperity distributions of metallic surfacês. *J. Appl. Phys.* **29**, No. 8 (1958).
6. DYSON, J. and HIRST, W., *Proc. Phys. Soc.* B67, 309 (1954).
7. KRAGELSKII, I. V. and SABELNIKOV., *Proc. Inst. Mech. Engrs* 302 (1967).
8. BOESCHTEN, F. and VAN DER HELD, E. F. M., *Physics* **23**, 37 (1957).
9. BOWDEN, F. P. and TABOR, D., The area of contact between stationary and moving surfaces. *Proc. Roy. Soc.* **A169**, 391 (1939).
10. D'YACHENKO *et al.*, The actual contact area between touching surfaces. *Russ. Inst. Mech. Engrs* (1963).
11. GREENWOOD, J. A. and WILLIAMSON, J. B. P., Contact of nominally flat surfaces. *Proc. Roy. Soc.* A295, 300 (1966).
12. KIMURA, Y., An analysis of distribution of contact points through the use of surface profiles. *J. Japan. Soc. lubric Engrs* 467 (1966).
13. IWAKI, A. and MORI, M., On the distribution of surface roughness when two surfaces are pressed together. *Bull. Japan. Soc. mech. Engrs* **1**, 329 (1958).
14. FENECH, H., The thermal conductance of metallic surfaces in contact. Ph.D. Thesis, M.I.T. (1959).
15. TSUKIZOE, T. and HISAKADO, T., On the mechanism of contact between metal surfaces. *Trans. Am. Soc. mech. Engrs* Part 1; **87D**, 666; Part 2, **90F**, 81.
16. GREENWOOD, J. A. and TRIPP, J. H., Elastic contact of rough spheres. *Burndy Res. Div. Rep.* 21 (1965).
17. GREENWOOD, J. A. and TRIPP, J. H., Contact of nominally flat rough surfaces, *Burndy Res. Div. Rep.* 66 (1968).
18. BLOK, H., *Proc. Roy. Soc.* A212, 480 (1952).
19. HALLIDAY, J. S., *Proc. Inst. mech. Engrs* **169**, 177 (1955).
20. WHITEHOUSE, D. J. and ARCHARD, J. F., The properties of random surfaces in contact. *Symp. Surface Mech.* Proc. A.S.M.E. (1969).
21. KRAGELSKII, I. V. and DEMKIN, N. B., *Wear.* 3, 170 (1960).
22. WILLIAMSON, J. B. P. and HUNT, R., Asperity persistence and real area of contact between rough surfaces. *Burndy Res. Rep.* (1970).
23. PULLEN, J. and WILLIAMSON, J. B. P., Plastic contact of rough spheres. *Burndy Res. Rep.* (1970).

24. Moore, A. J. W., *Proc. Roy. Soc.* **A195**, 231 (1948).
25. Jaeger, J. C., *Proc. Roy. Soc.* 56, 203 (1942).
26. Yip, F. C. and Venart, J. E. S., *Proc. Inst. mech. Engrs* **182**, Part 3K (1968).
27. Roth, A. and Inbar, A., An analysis of the vacuum sealing process between turned surfaces. *Vacuum* **18**, 309 (1968).
28. Mitchell, L. A. and Rowe, M. D., *Proc. Inst. mech. Engrs.* **182**, Part 3K (1968).
29. Hahne, H., *CIRP Ann.* **17**, 387 (1969).
30. Ber, A. and Yarnitzky, Y., *Microtecnic* **22**, 449.
31. Kragelskii, I. V., *Friction and Wear*, p. 8 Butterworths (1965).
32. Lifshiz, E. M., *Dokl. Akad. Nauk. SSSR* 100 (1955).
33. Rabinowicz, E., *Friction and Wear of Materials.* John Wiley (1965).
34. Fogg, B., *Proc. Inst. mech. Engrs.* **182**, Part 3K (1968).
35. Buckley, D. H., *The Influence of the Atomic Nature of Crystalline Materials on Friction.* ASLE–ASME (1967).
36. Buckley, D. H., *Effect of Recrystallization on Frictional Properties of some Metals in Single Crystal and Polycrystalline Forms.* D-4143, N.A.S.A., TN (1967).
37. Adam, N. K., *The Physics and Chemistry of Surfaces.* Clarendon Press (1938).
38. Peterson, M. G., Florek, J. J. and Lee, R. E., Sliding characteristics of metals at high temperatures. *Trans. Am. Soc. lubric. Engrs* **3**, 101.
39. Myers, N. O., Characterisation of surface roughness. *Wear* **5**, 182 (1962).
40. Trott, B. B., British Nylon Spinners (1963).
41. Archard, J. F., *Wear. NASA Symp.*, San Antonio (1969).
42. Hirst, W. and Lancaster, J. K., Surface film formation and metallic wear. *J. Appl. Phys.* **27**, 1057 (1956).
43. Burwell, J. T., A survey of possible wear mechanisms. *Wear* **1**, 119 (1957).
44. Rabinowicz, E. and Tabor, D., Metallic transfer between sliding metals. *Proc. Roy. Soc.* **A208**, 455 (1951).
45. Kerridge, M. and Lancaster, J. K., The stages in a process of severe metallic wear. *Proc. Roy. Soc.* **A236**, 250 (1956).
46. Kruschov, M. and Babichev, M. A., *Friction and Wear in Machining*, Vol. 19. ASME (1965).
47. Dawson, P. H., Elastohydrodynamic lubrication. *Proc. Inst. mech. Engrs* **180**, Part 3B, 95 (1966).
48. Waterhouse, R. B., Fretting corrosion. *Proc. Inst. mech. Engrs* **169**, 1157 (1955).
49. Cameron, A., The surface roughness of bearing surfaces and its relation to oil film thickness at breakdown. *Proc. Inst. mech. Engrs.* (1949).
50. Tallian *et al.*, Lubricant films in rolling contact of rough surfaces. *Trans. Am. Soc. lubric. Engrs.* (1964).
51. Furey, M. J., Metallic contact and friction between sliding surfaces. *Trans. Am. Soc. lubric. Engrs* (1961).
52. Christensen, H., Nature of metallic contact in mixed lubrication. *Proc. Inst. mech. Engrs* **179**, 100 (1965).
53. Imai, M. and Rabinowicz, E., *Trans. Am. Soc. lubric. Engrs* **6**, 286 (1963).
54. Van Hasselt, R., Conference on Manufacturing Technology. CIRP/ASTME, Ann Arbor (1967).
55. Littman, W. E., CIRP/ASTME, Ann Arbor (1967).
56. Field, M., Koster, W. P. and Kahles, J. F., CIRP/ASTME, Ann Arbor (1967).
57. Kasper, A. S., *Soc. Auto. Engrs* SP. 268 (1965).
58. Reneau, J. and Collinson, J. A., *B.S.T.J.* **44**, 2203.
59. Beckmann, P. and Spizzichino, A., *The Scattering of Electromagnetic Waves from Rough Surfaces.* Pergamon Press (1963).
60. Fraiture, L., Louvain, Wouters, p. 61 (1964).
61. Bennett, H. E. and Porteus, J. O., *J. opt. Soc. Am.* **51**, 123 (1961).
62. Davies, H. *Proc. Inst. elec. Engrs* **101**, 209 (1954).
63. Hasunuma, H. and Nara, J., On the sheen gloss. *J. Phys. Soc. Japan* **11**, 69 (1956).
64. Richmond, J. C. and Steward, J. B., *J. Am. ceram. Soc.* **42**, 633 (1959).
65. Richmond, J. C., *J. opt. Soc. Am.* 56, 253 (1966).
66. Skinner, —., *J. Opt. Soc. Am.* (1963).
67. Kahles, J. F. and Field, M., *Proc. Inst. Mech. Engrs* **182**, Part 3K (1968).
68. Sata, T., Surface finish in metal cutting. *CIRP Ann.* **12**, 190 (1963).
69. Hoshi, K. and Hoshi, T., On the metal cutting with built-up edge. *Proc. 9th MTDR Conf.*, p. 1099. Pergamon Press (1968).
70. Takeyama, H. and Ono, T., Study of roughness on turned surfaces. *Bull. Japan. Soc. precis. Engrs* **1**.
71. Shaw, M. C. and Sanghani, S. R., On the origin of cutting vibrations. CIRP (1962).
72. Von Turkovich, B. and Calvo, S., *Proc. 9th MTDR Conf.* p. 1051. Pergamon Press (1968).
73. Von Turkovich, B. and Micheletti, G. F., *Proc. 9th MTDR Conf.* p. 1073. Pergamon Press (1968).
74. Samuels, L. E. and Wallwork, G. R. *J. Iron Steel Inst.* **186**, 211 (1957).
75. Peklenik, J., Contributions to the theory of grinding. *Mech. J.* 4/5 (1959).
76. Kaliszer, H. and Singhal, P. D., Analysis of waviness produced through grinding. *CIRP Ann.* **15**, 245 (1967).
77. Farmer, D. A., Brecker, J. N. and Shaw, M. C., Study of the finish produced in surface grinding. *Proc. Inst. mech. Engrs* **182**, Part 3K, 171 (1968).
78. Barker, K., Kaliszer, H. and Rowe, G. W., *Proc. Inst. mech. Engrs* **182**, Part 3K, 195 (1968).
79. Wetton, A. G., A review of theories of metal removal in grinding. *J. Mech. Eng. Sci.* **11**, 412 (1969).
80. Samuels, L. E., *Metallographic Polishing by Mechanical Means.* Pitman (1967).
81. Bielby, G., *Aggregation and Flow of Solids.* MacMillan (1921).

82. Jobling, A. V., Forging principles and metallurgy. *Mod. Workshop Technol.* 331 (1966).
83. Olsen, K. V., Surface roughness on turned steel components and the relevant mathematical analysis. *Prod. Engr* 593 (1968).
84. Van Hasselt, R., The need for developing a typology of surfaces. *CIRP Am.* **15**, 349 (1967).
85. Reason, R. E., Le calcule automatique des criteres des profils de surfaces. *Automatisme* **9**, No. 5, 177 (1964).
86. Bendat, J. S., *Principles and Applications of Random Noise Theory*, p. 213. John Wiley (1958).
87. Papoulis, A., *Probability, Random Variables and Stochastic Processes*, p. 475. McGraw Hill (1965).
88. Bendat, J. S. and Piersol, A. G., *Measurement and Analysis of Random Data*. John Wiley (1966).
89. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra*. Dover (1958).
90. Krendel, E. S., *I.R.E. Trans. Med. Elect.* 149 (1959).
91. Storm van Leeuwen, W., *Elec. Clin. Neurophysiol.* **16**, 136 (1964).
92. Ward, J. F., *Nature* **223**, 1325 (1969).
93. Gumbel, E. J., *Statistics of Extremes*. Columbia University Press (1959).
94. Abbott, E. J. and Firestone, F. A., *Specifying Surface Quality*. Mech. Engng. (1933).
95. Pesante, M., Determination of surface roughness typology by means of amplitude density curves. *CIRP Ann.* **12**, 61 (1963).
96. Reason, R. E., The bearing parameters of surface topography. *Proc. 5th MTDR Conf.* Pergamon Press (1964).
97. Ehrenreich, M., *The Slope of the Bearing Area as a Measure of Surface Texture*. Microtecnic (1959).
98. Al-Salihi, T., Ph.D. Thesis, University of Birmingham (1967).
99. Von Weingraber, H., Accuracy and reliability of roughness measurements. CIRP (1969).
100. Reason, R. E., *Proc. Inst. mech. Engrs* **182**, Part 3K, 299 (1968).
101. Butler, R. D. and Pope, R. J., Surface roughness and lubrication in sheet steel metalworking. *Proc. Inst. mech. Engrs* **182**, Part 3K, 162 (1968).
102. Wormersley, J. R. and Hopkins, M. R., *J. Etats Surface* 135 (1945).
103. Linnik, Y., and Khusu, A. P., Mathematico—statistical description of surface profile irregularity in grinding. *Inzhernernyi, Sborn.* 20, 154 (1954).
104. Nakamura, T., *J.S.P.M.J.*, 25, 56 (1959); *J.S.P.M.J.*, 26, 226 (1960).
105. Peklenik, J., Investigation of the surface typology. *CIRP Ann.* **15**, 381 (1967).
106. Whitehouse, D. J. and Archard, J. F., to be published.
107. Hannan, E. J., *Time Series Analysis*. Methuen (1960).
108. Brooks, C. E. P. and Carruthers, N., *Handbook of Statistical Methods in Meteorology*. HMSO (1953).
109. Whitehouse, D. J. and Archard, J. F., The properties of random surfaces of significance in their contact. *Proc. Roy. Soc. Lond.* (1970).
110. Ber, A. and Braun, S., Spectral analysis of surface finish. *CIRP Ann.* **16**, 53 (1968).
111. Dunin-Barkovsky. Analysis of surface irregularities by the spectral method. *Proc. Inst. mech. Engrs* **182**, Part 3K, 211 (1958).
112. Myers, N. O., Characterisation of surface roughness. *Wear* **5**, 182 (1962).
113. Peklenik, J., Contribution to the theory of surface characterisation. *CIRP Ann.* **12**, 173 (1963).
114. Cooley, J. W. and Tukey, J. W., *Math. Comput.* 19, 297 (1965).
115. Kubo, M., *Rev. Sci. Inst.* 36, 236 (1965).
116. Nara, J., Some analysis of paper finished surfaces. *J.S.P.M.J.* 28, 120 (1962).
117. Spragg, R. C. and Whitehouse, D. J., A new unified approach to surface metrology. Inst. Mech. Engrs. (1970).
118. Longuet-Higgins, M. S., Statistical analysis of a random moving surface. *Proc. Roy. Sor.* **249A**, 966, 321 (1957).
119. Strahler, A. N., Hypometric analysis of erosional topography. *Geol. Soc. Am. Bull.* **63**, 1117 (1952).
120. McAdams, H. T., Picciano, L. A. and Reese, P. A., A computer method for hypsometric analysis of abraded surfaces. *Proc. 9th MTDR Conf.* p. 73. Pergamon Press (1968).
121. Peklenik, J. and Kubo, M., A basic study of a three-dimensional assessment of the surface generated in a manufacturing surface. *CIRP Ann.* **16**, 257 (1968).
122. Kubo, M. and Peklenik, J., An analysis of microgeometric isotropy for random surface structures. *CIRP Ann.* **16**, 235 (1968).
123. Ostvik, R. and Christensen, H., Changes in surface topography with running-in. *Tribology Convention*, Gothenburg, p. 59 (1969).
124. Williamson, J. B. P., Topography of solid surfaces. *NASA Symp.* (1970).
125. Hunt, R., Relocation profilometry. *J. Sci. Instrum.* **1**, (1968).
126. Krendel, E. S., *I.R.E. Trans. Med. Elect.* 149 (1959).
127. Liu, S. C., *Bell Syst. Tech. J.* 2273 (1968).

PROGRAM PROF FLOW CHART



In what follows some details will be given of the computer programs that have been used in the work described in this thesis. All of the programs have been written by the author; programs CUPE, PECU, REST and GRIN especially for this thesis, and programs PROF and ROLL modified to take into account work carried out in this thesis. All these modifications have also been written by the author.

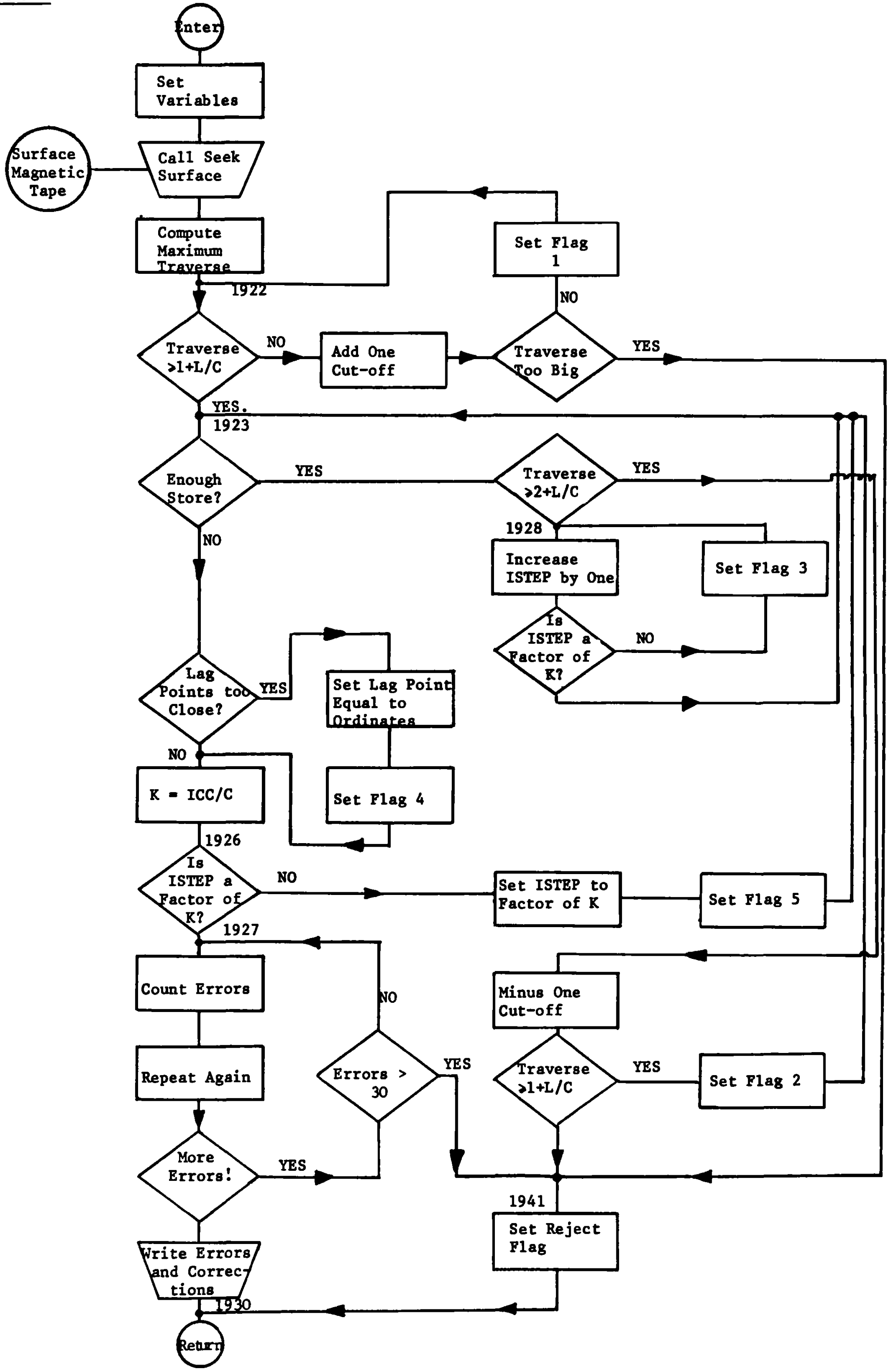
The programs will be limited simply to a flow chart and a short description including input variables, and they will be presented in an abbreviated form for clarity.

It is hoped that these lists of input variables, the short description, and the flow charts will give some idea of what these programs do. It obviously is not sufficient to allow a complete understanding but in view of the great length of many of them, which makes it impractical to include them all complete, it is felt that enough has been included to allow the references made to them in the thesis to be intelligble.

Program PROF

This program computes the basic statistical parameters of the profile. Basically in operation the profile itself can be processed, prior to statistical evaluation, by means of both high and low cut filters or simple DC removal. After filtering the following statistical parameters can be measured:

SUBROUTINE CHECKS
FLOW CHART



1. Ordinate height distribution with moments.

2. Peak and valley distributions with average peak and valley separations and average spacings.

3. Slope distribution with moments, either smoothed or not.

4. Average wavelength.

5. Autocorrelation function.

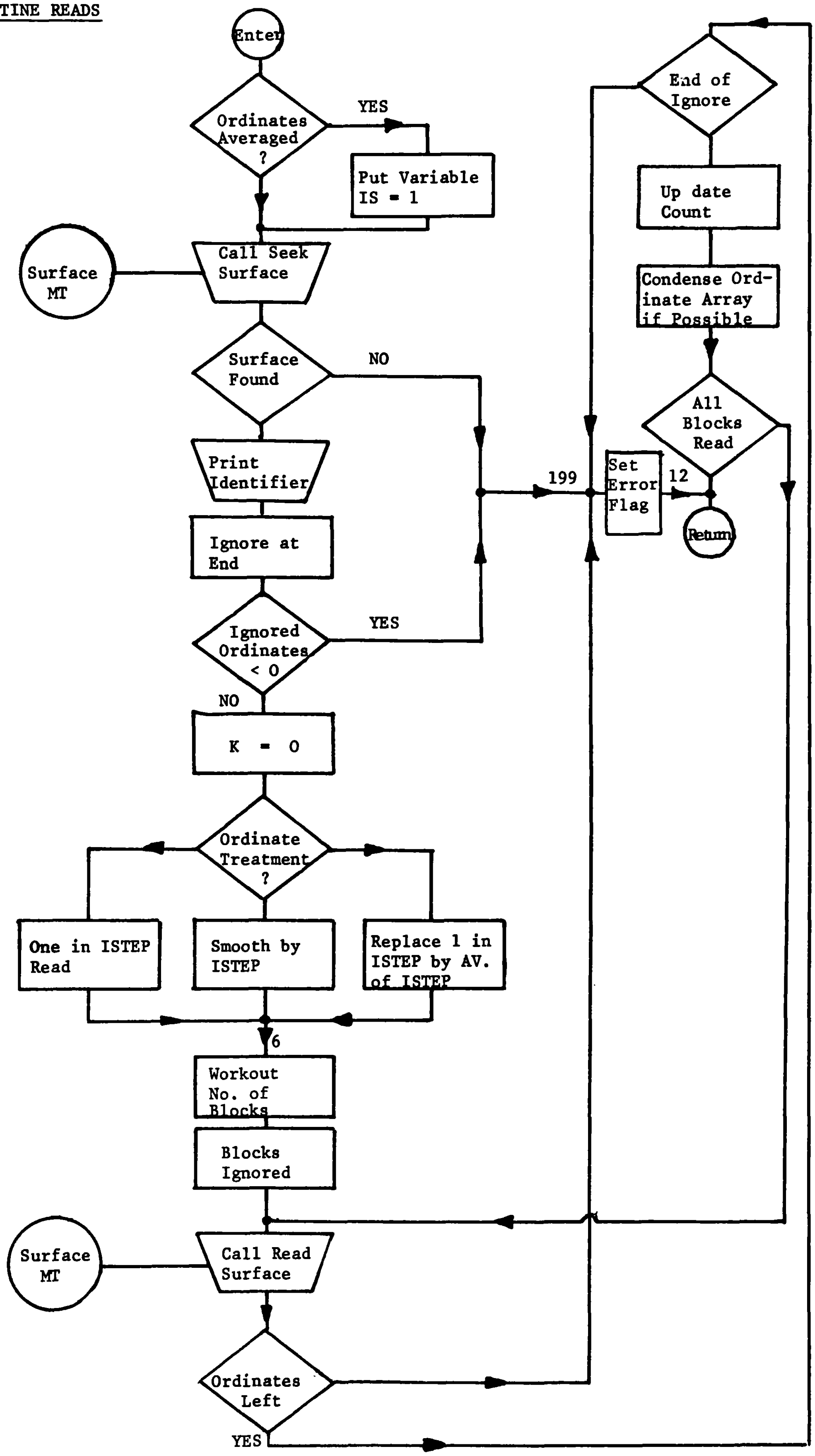
6. Structure function.

7. Power spectrum.

A number of options are open in deciding the particular computing path taken through the program. These will become apparent when the input variables are discussed. Subroutines READS and CHECKS are used to read in and check control variables and surface data; flow charts are provided for these. Subroutine SKETCH is used for plotting results.

To enable a condensation of the number of input variables needed the program automatically scales the plotted curve to fit best the available space. Also, in the mode of operation where both the profile and modified profile can be outputed they are made to share the plot array. Hence they are displayed alongside each other, which enables a direct comparison of shapes to be made.

SUBROUTINE READS



For convenience all the input variables are coded; the code number preceding the variable on the card. The list of variables is given overleaf.

INPUT VARIABLES (typical numerical values in brackets)

(a) Real numbers

Code	Variable	Assumed Value if Omitted
1.	Traverse length of profile measured in units of the filter cut-off (TRAV). (5.0)	Must be present
2.	Filter cut-off in mm (0.8)))	If not present then bandwidth ratio must be made to the smoothing filter cut-off; then only smoothing filter used.
3.	Filter cut-off in inches (0.03))	
4.	Profile ordinate psacing in) micrometres (PROFSP). (2.5)))	One must be present
5.	Profile ordinate spacing in) inches (PROFSP). (.0001))	
6.	Amount of data to be ignored at end of traverse to allow synchronism with meter (IGNORE). (0.0)	Assumed value zero
7.	Maximum autocorrelation lag in terms of the cut-off of the low cut filter (DELMAX) (0.05)	VALUE = 0.1 (TRAV - weighting function length).
8.	Slope of phase-corrected filter (B)	VALUE = 0.333
9.	Bandwidth ratio (BW) (100)	No top cut filter used
10.	Vertical magnification (VMAG) (10 000)	Must be present

(b) Integer numbers

Code	Variable	Assumed Value if Omitted
11.	Number of weighting function ordinates per cut-off (C) (50)	VALUE = 100
12.	Type of filter to be used (JOHN)	
	1. DC level removed	
	2. Standard 2 CR filter	VALUE = 2
	3. Phase-corrected filter	
14.	Length of weighting function in terms of ordinates (L) (100)	VALUE = 2XC
15.	Number of lag positions per cut-off in correlation function (IDD) (150)	No correlation function, structure function or power spectrum
16.	Determines the ratio of ordinates read to ordinates actually used (ISTEP) (1)	VALUE = 1
17.	The surface number on magnetic tape.	

The order of these variables must be real then integer and finally magnetic tape name. The order within these groups is not important providing that the vertical magnification is the last one in the real group and the surface number the last of the integer group.

How the ISTEP facility is used depends on the variable LIK which is obtained from a coding of the last two figures on the vertical magnification card, for example if VMAG is 10000.02 then

VMAG is taken as 10000 and LIK as 2.

For LIK equal to unity then the program reads only one in ISTEP ordinates.

For LIK equal to two then the program smoothes all the profile data over ISTEP ordinates.

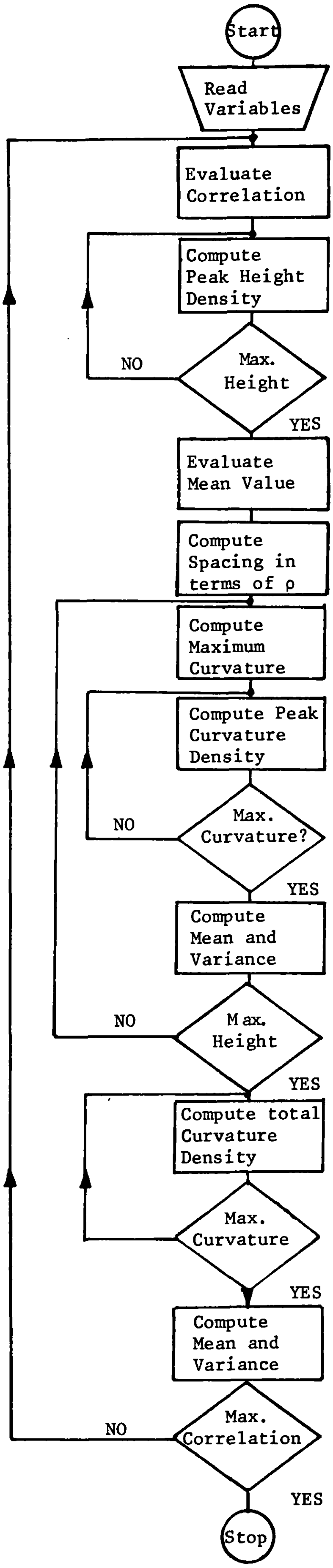
For LIK equal to three the program replaces every ISTEP ordinate by their average and reads just the average values in.

Another facility called ITEST is also included which is not relevant to the normal use. In normal mode up to ten surfaces can be computed for one run. This enables the card reader to be released for other users in the meantime. If less are to be run then the control variables for the last surface should be followed by a card containing an integer zero followed by at least one space then 0.0.

This and the other programs are written in ICL FORTRAN IV. The approximate run time for a surface is about ten minutes. The core requirement is for about 26 000 locations of immediate access store and magnetic tape transports containing the surface profile data.

To print as plotted lines the mean line and modified profile complete then activating switch 2 is necessary. This is simply achieved by instructing the operator to type ON ⨳ PROF 2. Activating switch 1 has the effect of simply bypassing the autocorrelation function, structure function and power spectrum.

PROGRAM PECU
FLOW CHART


PEAK DISTRIBUTION
VARIATION OF PEAK CURVATURE WITH HEIGHT
TOTAL CURVATURE
Start
Read Variables
Evaluate Correlation
Compute Peak Height Density
Max. Height
NO
YES
Evaluate Mean Value
Compute Spacing in terms of ρ
Compute Maximum Curvature
Compute Peak Curvature Density
Max. Curvature?
NO
YES
Compute Mean and Variance
Max. Height
NO
YES
Compute total Curvature Density
Max. Curvature
YES
Compute Mean and Variance
Max. Correlation
NO
YES
Stop

Program PECU

This program computes the numerical solutions of the equations derived in Chapter 4 for peak height distributions, peak curvature at different heights and the various moments of the distributions. The output is printed on the line printer in a crude plot form. Obviously, because of the symmetry of the random waveforms, valleys can be evaluated from the same output.

It is possible to get solutions to the equations for a wide range of value of correlation.

The relevant input variables with typical values in brackets are:

EXTR (3.5) The extreme height value to be used.

STEP (0.5) The increment in height values.

RATIO (30.0) The ratio of the correlation length to that of the standard deviation.

ROL (0.9) The last correlation value to be considered.

RINC (0.1) The increment of correlation.

ROS (0.1) The first correlation value.

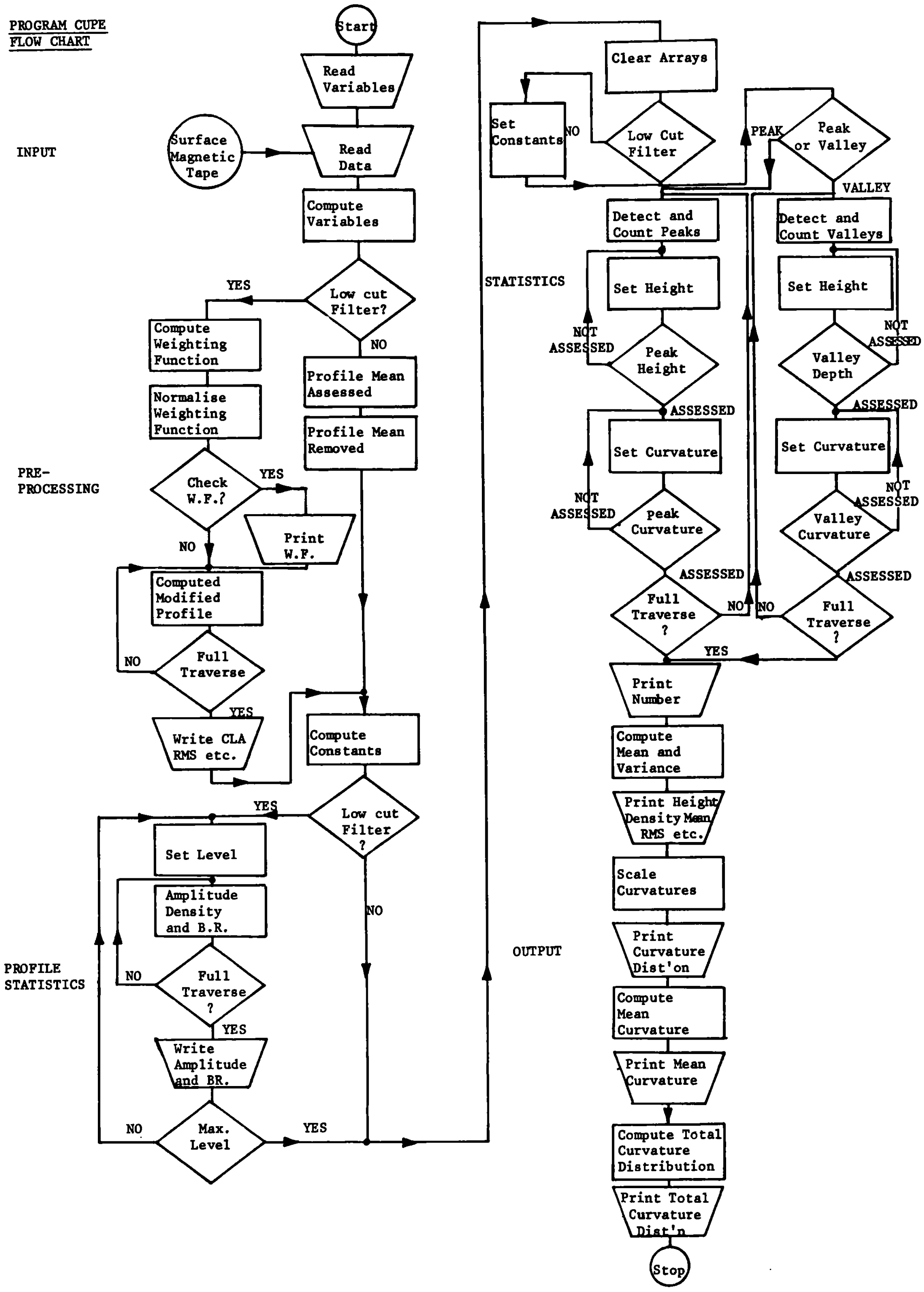
ROLV 25 The value of the autocorrelation assumed for independence.

NOCURV 5 The number of increments required in the curvature distribution.

SIZE 1, 2, 5 (5) Scaling factors for the plot routine.

This program takes a few minutes to run and is 4914 store locations in size.

PROGRAM CUPE
FLOW CHART



Program CUPE

This program measures the peak height distribution and peak curvature distributions as a function of height using three-point analysis for practical surfaces. It, therefore, uses the magnetic tape library of surface profiles.

In operation, first the data, picked up from magnetic tape, is processed, that is any irrelevant low frequency characteristics are removed by means of the phase-corrected filter. Second the amplitude distribution and bearing ratio are computed. This gives the RMS value which is used as a unit. After this the required peak height (or valley) distributions and curvatures are evaluated as indicated on the flow chart. One feature of the output, which is again in the crude plot form, is that it is made to be exactly similar in format to PECU so that an immediate comparison between theoretical and practical results could be made at a glance, even to comparing the shape of the distributions.

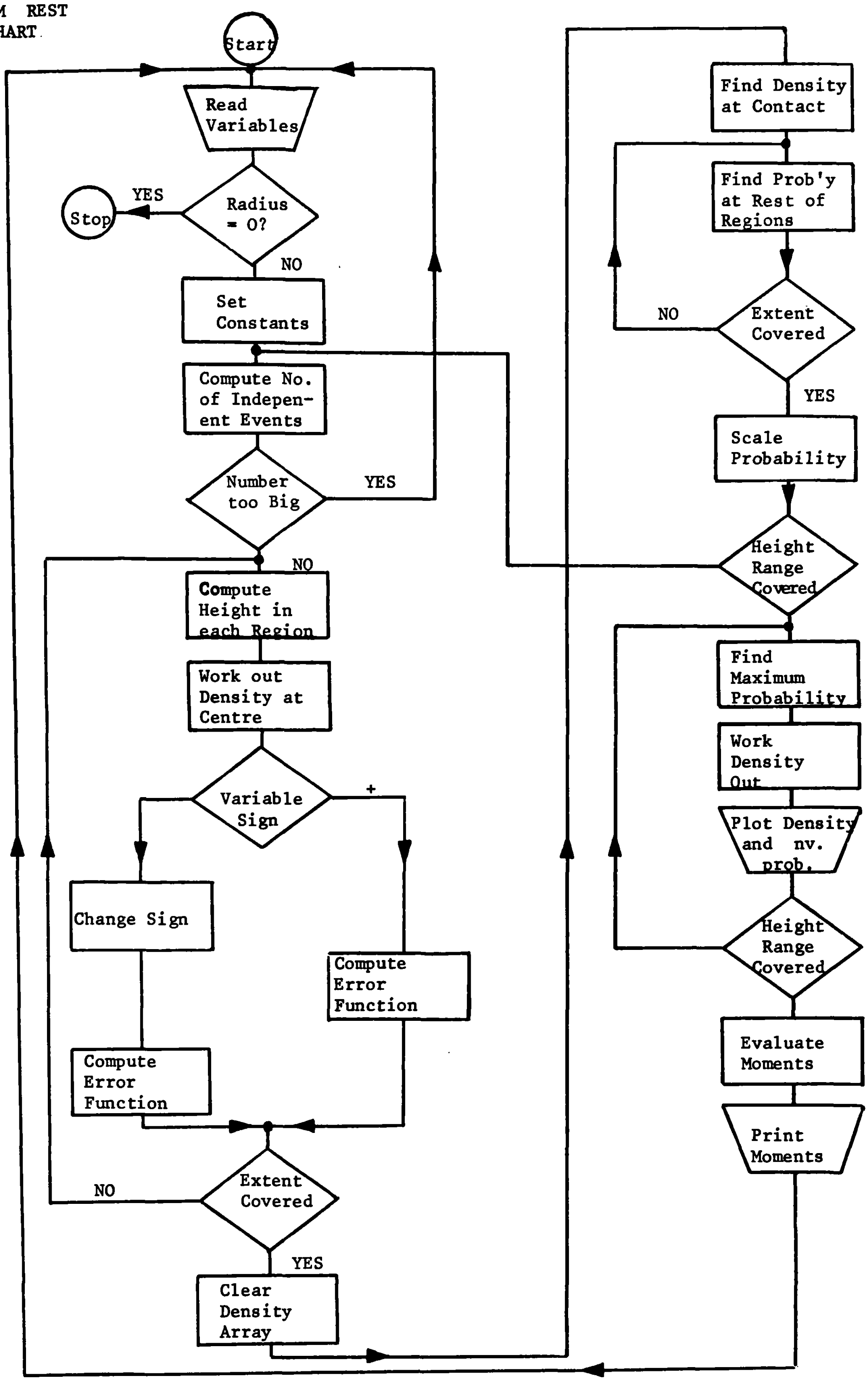
Input variables with typical values in brackets

TRAV	(5.0)	Assessment length in filter cut-off values.
END	(0.0)	Length of surface to be ignored at the end.
SIZE 1,2,3,4,5		Scaling factors for the plot routine. Typical values taken in order are 100.0, 0.5, 300.0, 200.0, 200.0.
SPACE	(40.0)	Ordinate spacing in micro-inches.
VMAG	(10000.0)	Vertical magnification.

ROL	(600)	Distance of independence in micro-inches.
ROLV	(0.10)	Correlation value taken as independence.
CURVV	(2.5)	Maximum likely value of curvature measured in terms of the standard deviation of the surface.
RFRAC	(0.5)	Increment of height in terms of the standard deviation of the surface.
VALUE	(20.0)	RMS value in micro-inches for the non-filtered case.
C	(50.0)	Number of weighting factors in one cut-off length of the filter.
JOHN	(1)	Equals unity then weighting function printed, if equal to two then not printed.
JIM	(1)	If equal to unity filters the profile and gives amplitude distribution.
NOCURV	(25.0)	The number of points required in the curvature distribution.
ICC	(150)	The number of profile ordinates in the cut-off of the filter.
L	(100)	The length of the weighting function in ordinate spacings.
ISKIP	(1)	The number of ordinates missed out on either side of the central ordinate.
ISTEP	(1)	The number of points taken as the central ordinate compared with the total number of profile ordinates available.

The program takes about ten minutes to execute and uses 23,329 store locations. The maximum number of ordinates to be evaluated at one run is four thousand five hundred.

PROGRAM REST
FLOW CHART



Program REST

The numerical solution of the equations developed in Section 5.2.2 are evaluated in program REST. In it a circular body is imagined to be run across a random surface. The program compares the probability density of the locus of the envelope movement with the amplitude probability density of the profile itself. Both curves are plotted on the same scale and on the same axis to enable an immediate comparison to be made. In addition to the curves the mean value and RMS values of the distributions are found. One other feature of the output routine is that the graphs are automatically scaled to best fit the available space.

Input variables.

RAD   The radius of the circular part in terms of the distance of independence of the random surface.

RATIO   The ratio of the independence distance to the RMS value.

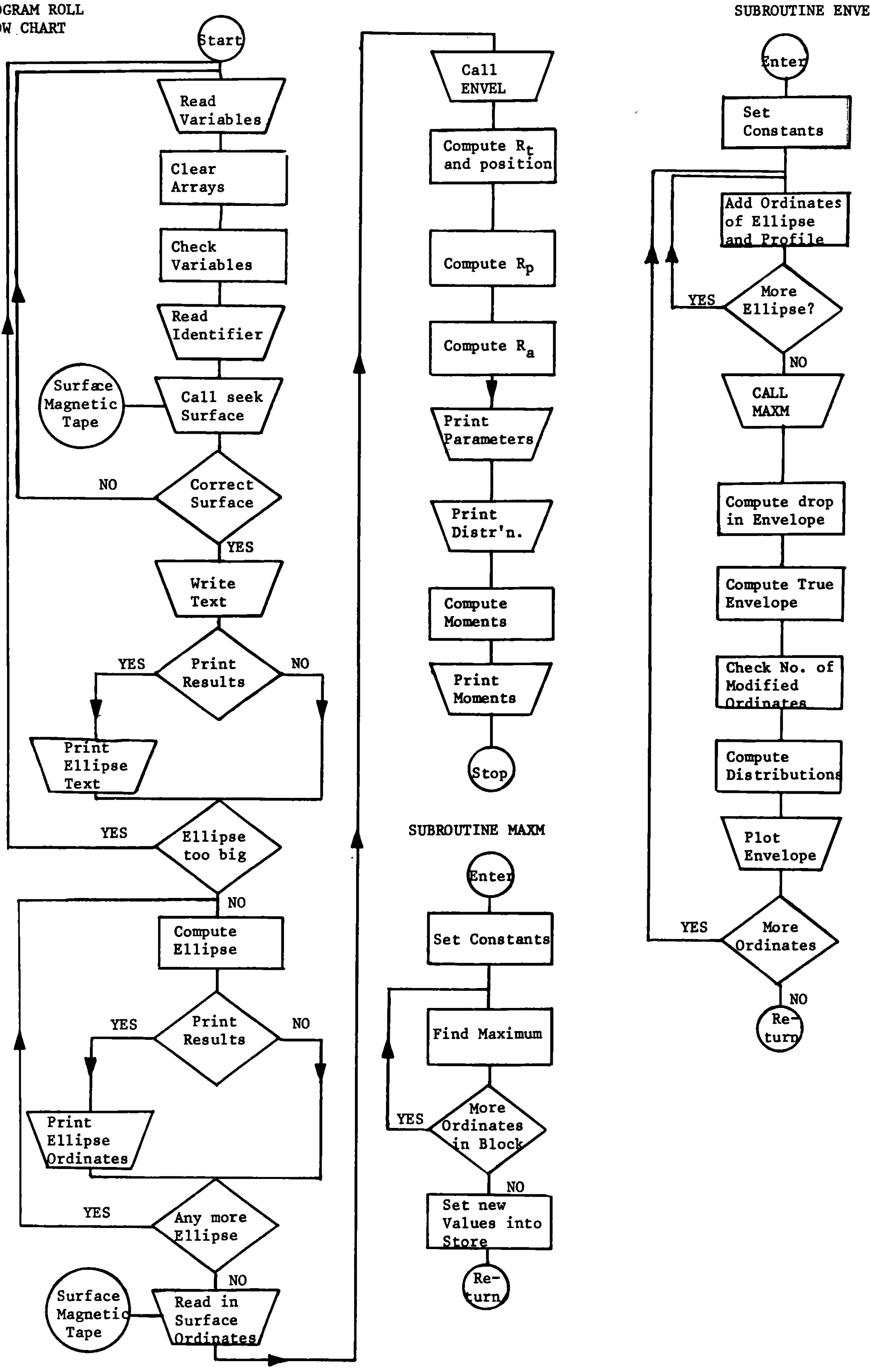
STEP   The increment of height to be considered in the distribution.

This program takes a few minutes to run and requires 4,544 store locations.

PROGRAM ROLL
FLOW CHART

SUBROUTINE ENVEL

SUBROUTINE MAXM



Program ROLL

As with PECU and CUPE, ROLL is the practical equivalent of the program REST in which analytic formulae are evaluated. In ROLL the ordinates are read in block at a time and a circle (modified to an ellipse because of the magnification distortion) is positioned with its centre above each ordinate in turn. The point of contact of the circle with the profile is then found for each position according to equation 3-38. Having found this the envelope position can be easily deduced. This process is carried out over three thousand ordinates. The area enclosed between the envelope and the profile is worked out at the same time as the relative frequency of the profile height and envelope height. This is done in Subroutine ENVEL which also plots out the profile and envelope. From the area the R_p value (or average depth of profile from the envelope is worked out and also the maximum departure R_t. Then the envelope is dropped by a value of R_p and the R_a (or average departure from the envelope and profile worked out. Finally the distributions of the envelope and profile are plotted together with their moments.

Input variables.

J	(500)	The number of ordinates in a block
IEND	(0.0)	The number of ordinates left at the end of the traverse to enable a direct tie up between this and the other methods of assessment.

ITRAV (2000) The number of profile ordinates to be assessed.

H (100.0) Horizontal magnification.

V (5000.0) Vertical magnification.

R (2.0) Radius of circle in inches.

CWFACT (0.25) Factor in inches of chart width used to estimate extent of ellipse to be used.

PROFSP (.0001) Profile ordinate spacing in inches.

IOPT (3) Determines mode of output

VALUE = 1 Ellipse ordinates printed together with result.

= 2 Envelope and profile printed together with results.

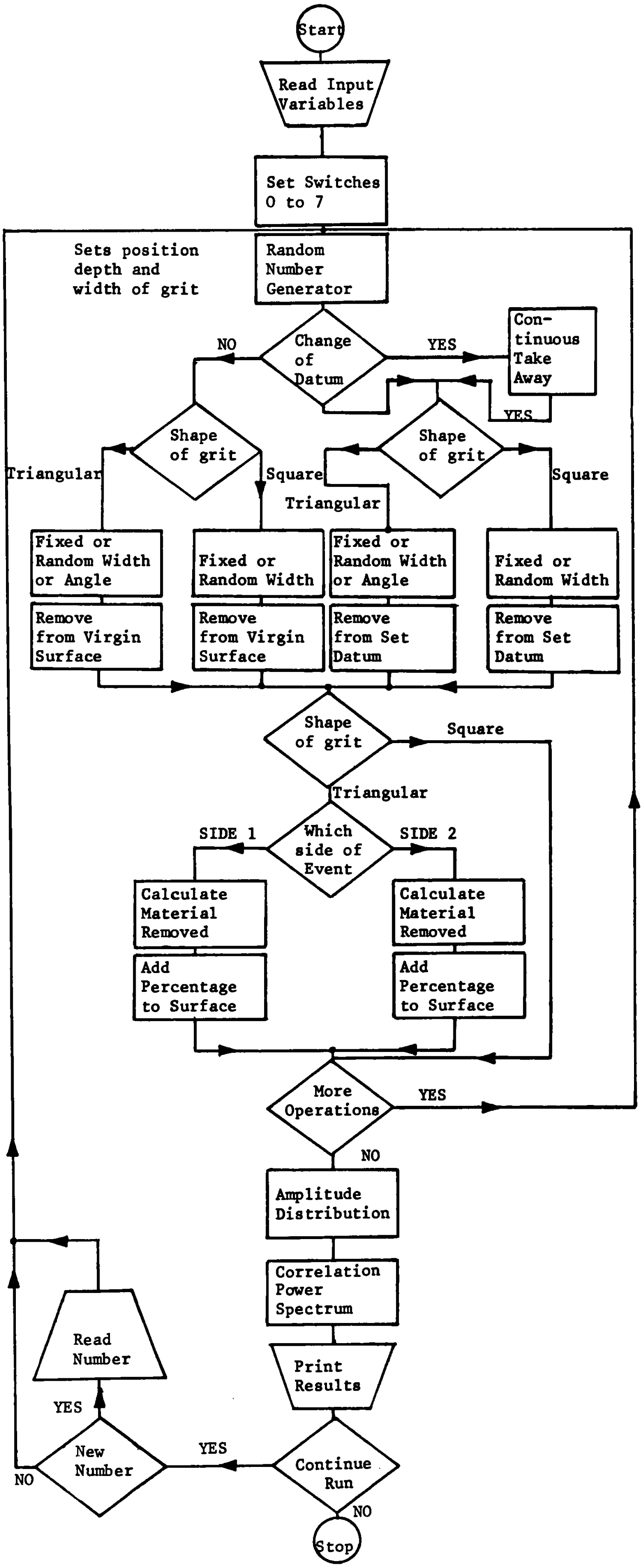
= 3 All results outputed.

= 4 Only final results outputed.

INO (1) For INO = n then every n^{th} point on the envelope and profile plotted.

This program takes about 15 minutes to run normally and uses a store of 20,000 locations.

PROGRAM GRIN
FLOW CHART



Program GRIN

This program simulates a random machining process. It first generates a random number of eight figures which is broken down to give (a) the position where the "grit" hits the surface, (b) the depth to which it penetrates, and (c) the width of the grit.

It is possible to use the program in a number of modes ranging from the use of square grit to that of triangular shapes of various kinds. The probability distribution of the grit heights is arbitrary and and is set by means of an input card. Also the size of the grit can be fixed or variable. In one important mode the amount of ploughing produced by the grit (and hence something akin to the hardness of the material) can be simulated.

After a period of "machining" there is a facility in this program for the statistics of the generated profile to be measured after which the machining is continued. The statistical parameters that are measured are the profile height distribution with moments, the autocorrelation function and the power spectrum. After each such measurement the resultant profile and statistical parameters are listed using the plot routine used extensively in all these programs.

Switch facilities

Switch 0	When on a fixed width of square grit is used (Switch 4 has to be on also).
Switch 1	When on a print out of the surface profile is given. This switch controls variable JIM.

Switch 2 When on a new random number primer will be read. This switch controls the variable JAMES.

Switch 3 When on a new origin for the grit height distribution is requested. This shift is called DROP in the program and is controlled by the variable JED.

Switch 4 When on a rectangular (or square) grit is used. The variable is JEF.

Switch 5 When on a triangular grit of 45° angle is used. The variable used is FIXANG.

Switch 6 When on the one side of the triangular grit is suppressed. The variable is ONEANG.

Switch 7 When on the variable HARD is read in. This gives the percentage of metal removed that has to be displaced. The variable used is JEN.

Input variables via the card reader.

IT (5) Only used when switches 0 and 4 on. This determines the half width of the fixed rectangular grit.

ANR (0.1672401) The primer for the random number generator.

IBLO (20 values) This is a card containing twenty numbers describing a distribution of grit heights. This is only brought in if ANR is made negative, otherwise a uniform height distribution is assumed.

ICHECK (1500) The number of operations before the statistics are examined.

ITOT (3) The number of runs

LIMIT (20) Tangent of the maximum angle of the grit.

JOHN (2) If equal to 1 a continuous take away of material from no fixed base line is assumed. If equal to 2 then the drop facility can be used.

The optional variables which follow the above are:

BOTANG (0.0) Only read when switches 4 and 5 are off. It is the tangent of the lowest angle of grit.

HARD (4.0) Only read when switch 7 is on. It gives the amount of ploughing i.e. for 50% ploughing HARD = 2.0.

DROP (5.0) Only read in when switch 3 is ON.

The program takes about five minutes to make one run and uses 16094 store locations. The author gratefully acknowledges the assistance of Mr. G. Burger in tidying up the ploughing routine.

Proc. Roy. Soc. Lond A. **316**, 97–121 (1970)
Printed in Great Britain

The properties of random surfaces of significance in their contact

By D. J. Whitehouse
Rank Precision Industries Ltd

and J. F. Archard
Department of Engineering, The University of Leicester

(*Communicated by D. Tabor, F.R.S.—Received* 11 *August* 1969—
Revised 27 *October* 1969)

In recent work it has been shown that many types of surfaces used in engineering practice have a random structure. The paper takes, as a representation of the profile of such a surface, the waveform of a random signal; this is completely defined by two parameters, a height distribution and an auto-correlation function. It is shown how such a representation can be transformed into a model, appropriate for the study of surface contact, consisting of an array of asperities having a statistical distribution of both heights and curvatures. This theory is compared with the results of an analysis of surface profiles presented in digital form. The significance of these findings for the theory of surface contact and for the measurement and characterization of surface finish is discussed.

1. Introduction

All surfaces are rough. This is the starting-point from which current ideas about friction, wear, and other aspects of surfaces in contact have evolved. Because surfaces are rough the true area of contact, which is much smaller than the apparent area in contact, must support pressures so large that they are comparable with the strengths of the materials of the contacting bodies. In their earlier work Bowden & Tabor (1954) suggested that these contact pressures are equal to the flow pressure of the softer of the two contacting materials and the normal load is then supported by plastic flow of its asperities. The true area of contact, A, is then proportional to the load, W; thus it was possible to provide a simple and elegant explanation of Amontons's laws of friction. However, if the asperities are plastically deformed the details of the surface finish seem relatively unimportant since the total area of contact and the contact pressure do not depend upon surface topography.

More recently it has been recognized that surface contact must often involve an appreciable proportion of asperity contacts at which the deformation is entirely elastic. It has been shown that, under conditions of multiple contacts, even if the deformation were entirely elastic, the true area of contact, A, can increase almost proportionally with the load, W (Archard 1957; Greenwood & Williamson 1966); thus a satisfactory explanation of Amontons's laws of friction is not dependent upon the assumption of plastic deformation. It therefore becomes more important to understand, in some detail, the role which surface topography plays in the contact of surfaces. For example, it is clear that surface finish will play a large part in determining the proportions of elastic and plastic deformation which will occur under any given set of conditions.

Knowledge of the topography of surfaces has been derived from the use of many techniques of surface examination. However, in considering the contact of nominally flat surfaces, the most relevant information has come from the use of profile meters in which a lightly loaded stylus is moved across the surface. In the past it has sometimes been assumed that the only significant information thus obtained could be expressed as the r.m.s. or centre line average (c.l.a.) value of the profile. However, in more recent years the outputs of profile meters have been analysed in greater detail by both analogue and digital techniques. In the field of production engineering this information has been presented in many different ways; height distributions, slope distributions, power spectral density curves, and auto-correlation functions are but a few of the characteristics which have been displayed (Peklenik 1967–8).

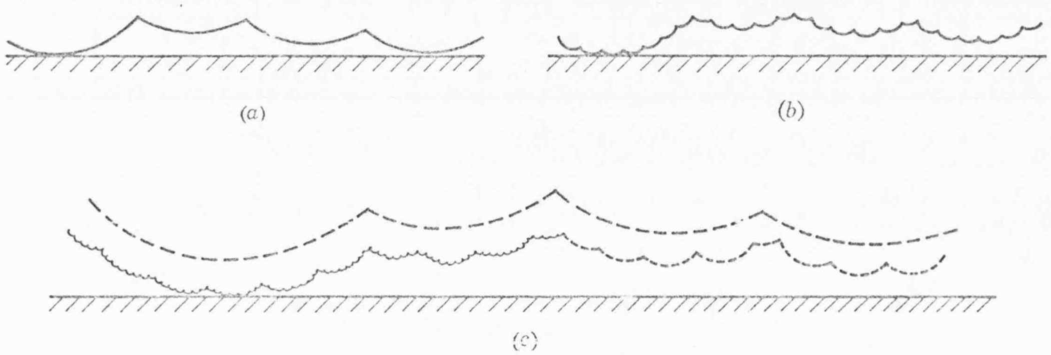


FIGURE 1. Models of surfaces containing asperities of differing scales of size. When the deformation is elastic the relationships between the area of contact (A) and the load (W) are as follows: (a) $A \propto W^{\frac{4}{5}}$; ($b$) $A \propto W^{\frac{14}{15}}$; ($c$) $A \propto W^{\frac{44}{45}}$.

On the other hand, those concerned with the problems of surface contact have used models of surfaces based upon many different assumptions about the nature of surface topography. Thus Archard (1957) postulated a series of models (figure 1) which were used to provide the first explanation of Amontons's laws of friction for elastic deformation of asperities. Although it was admitted that these models were artificial, they contain an important feature to which we shall revert later; this is the assumption that there exists upon the surfaces superposed asperities of widely differing scales of size.

It is, of course, important that the models used in theories of surface contact be more closely based upon knowledge gained from the examination of surface topography. Greenwood & Williamson (1966), and others, using information obtained by digital analysis of profile meter outputs, have shown that for many surfaces the distribution of heights is very close to Gaussian. Greenwood & Williamson also made an investigation of the height distribution of peaks; the most common technique used in this investigation was three point analysis, a peak being defined when the central of three successive sampled heights lies above those on either side. Thus the distribution of peak heights was also shown to be close to Gaussian but both the mean value and the standard deviation of this distribution differed from that of the heights of the ordinates. In addition, by the same techniques, a distribution of peak curvatures was obtained; this was skewed towards lower values of curvature. As a result of these observations Greenwood & Williamson postulated a model, representing a rough nominally flat surface, consisting of a series of spherical peaks, each having the same radius of curvature, and having a Gaussian distribution of heights. On the basis of this model, and the assumption that the deformation was elastic, it was shown that the relation between A and W was close to direct proportionality; thus a second theoretical derivation of Amontons's laws for elastic deformation conditions was provided.

The theory of Greenwood & Williamson (1966), although representing a notable advance, is still far from a complete or accurate representation of random surfaces such as those analysed in their work; in particular the assumption of a single radius of curvature for surface asperities is clearly a major simplification of the model. Moreover, it will be shown below that the use, in their examination of surface profiles, of three point analysis together with a single sampling interval severely limits the information obtained from the surface profile.

The present paper considers the representation of a surface profile as a random signal. Although such a representation can be a complete description of the profile the problem lies in its transformation into a form appropriate to the study of surface contact. From earlier work this requirement is seen as a model consisting of a distribution of asperities; therefore the paper is concerned with the distribution of the heights of the peaks and their radii of curvature.

2. The model

A surface profile (figure 2), if it is of a random type, can be defined completely (in a statistical rather than a deterministic sense) by two characteristics: the height distribution and the autocorrelation function (see, for example, Bendat 1958). In the main body of this paper we shall confine our attention to the particular example of a surface profile having a Gaussian distribution of heights and an exponential autocorrelation function. There are a number of reasons for using this particular model. First, and foremost, analysis by one of us (D. J. Whitehouse, to be published) of a large number of surfaces used in engineering practice has shown that a not insignificant proportion of such surfaces fits this model or is a reasonable approximation to it. Moreover, this model has been widely used in the theory of random processes; it has also been used to represent surfaces in studies of the scattering of electromagnetic radiation (Beckmann & Spizzichino 1963).

The model also simplifies some aspects of the mathematics and allows a clearer statement of the important physical principles to emerge. Throughout this analysis it will be assumed that the surfaces are isotropic although it is possible, at least in principle, to extend the theory to surfaces having an anisotropic structure.

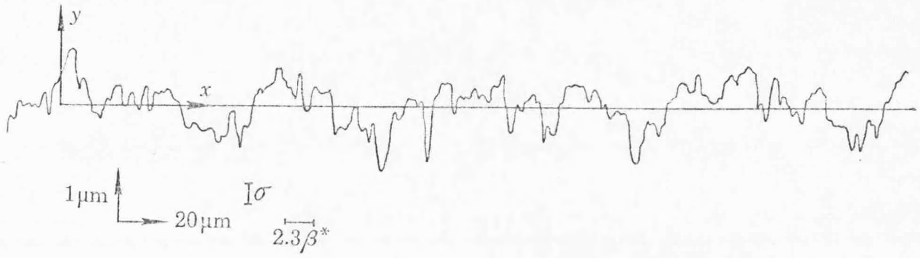


FIGURE 2. Surface profiles of Aachen 64–13 showing coordinate system. The profile is of a surface chosen for a detailed analysis which is described later in the paper. The magnitude of the r.m.s. value of the height distribution (σ) and the correlation distance β^* are shown for comparison with the profile.

The system of coordinates used is shown in figure 2; the mean line through the profile will be taken as $y = 0$. In practice the d.c. level, the general slope and the curvature of the surface are removed by a filter eliminating the longest wavelengths. This does not substantially affect the autocorrelation function. The probability of finding an ordinate at a height between h and $(h+\mathrm{d}h)$ is $f(h)\,\mathrm{d}h$. When the height distribution is Gaussian the height probability density function is

$$f(y) = (2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}y^2\right); \tag{1}$$

the heights h have now been expressed in the normalized form $y = h/\sigma$, where σ is the standard deviation of the height distribution.

The autocorrelation function of the profile is defined as

$$C(\beta) = \lim_{L\to\infty} \frac{1}{L} \int_{-\frac{1}{2}L}^{+\frac{1}{2}L} y(x)\, y(x+\beta)\, \mathrm{d}x, \tag{2}$$

where $y(x)$ is the height of the profile at a given coordinate x and $y(x+\beta)$ is the height at an adjacent coordinate $(x+\beta)$. In the theory which follows it will be assumed that

$$C(\beta) = \exp\left(-\beta/\beta^*\right), \tag{3}$$

where β^* will be called the correlation distance.† When $\beta = 2.3\beta^*$, $C(\beta)$ has declined to 10%; in what follows we shall, arbitrarily, take this spacing as that at which the two points on the profile have just reached the conditions where they can be regarded as independent events.

† We use the term 'correlation distance' to mark a distinction between ourselves and Peklenik (1967–8) who uses the term 'correlation length' for $2.3\beta^*$.

There exists a Fourier transform relation between the autocorrelation function $C(\beta)$ and the power spectral density function $P(\omega)$ of a random waveform given by

$$C(\beta) = \left(\frac{1}{2\pi}\right)\int_{-\infty}^{+\infty} P(\omega)\exp(\mathrm{i}\omega\beta)\,\mathrm{d}\omega. \tag{4}$$

For an exponential autocorrelation function the power spectral density function is represented by white noise limited only in the upper frequencies by a cut-off of 6 dB per octave. This is illustrated in figure 3. Thus the physical meaning of our model is that the main components of the surface profile consists of a band covering the lower frequencies (longer wavelengths). Shorter wavelength components exist but their magnitude declines with increasing frequency so that, in this range, the amplitude is proportional to wavelength. Therefore, in broad terms, the random signal representation of a profile which has been introduced here has features akin to superposed asperities of differing scales of size; it introduces the multiple scale of size features of figure 1*c* (cf. figure 1*a*) into a random model representation.

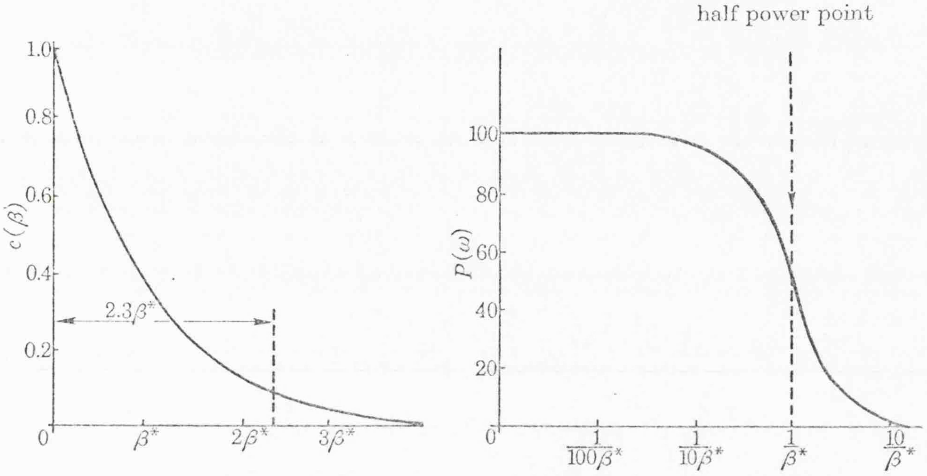


FIGURE 3. The model; autocorrelation function and power spectral density.

We are concerned here with the properties of surfaces of significance in their contact. These are the heights of the peaks and their curvatures; they will be defined, as in the earlier work, by three point analysis. This technique will first be applied to the theoretical model of our random profile and the results of this theory compared with experimental results obtained from the digital presentation of profilometer records.

3. THEORY

As a starting point in the development of the theory we assume that the profile has been sampled as a sequence of effectively independent events; therefore the ordinates considered are separated by lengths, $l \geqslant 2.3\beta^*$. Thus a peak at a height between y and $y+\delta y$ may be defined by three such consecutive events shown in figure 4a with the following restrictions: (a) the central event lies between y and $y+\delta y$; (b) event (1) has a value of less than y; (c) event (3) also has a value of less than y. Thus the probability that the central ordinate represents a peak between y and $y+\delta y$ is the multiplication of P_1, P_2 and P_3 where the P's refer to the shaded areas of the height distributions.

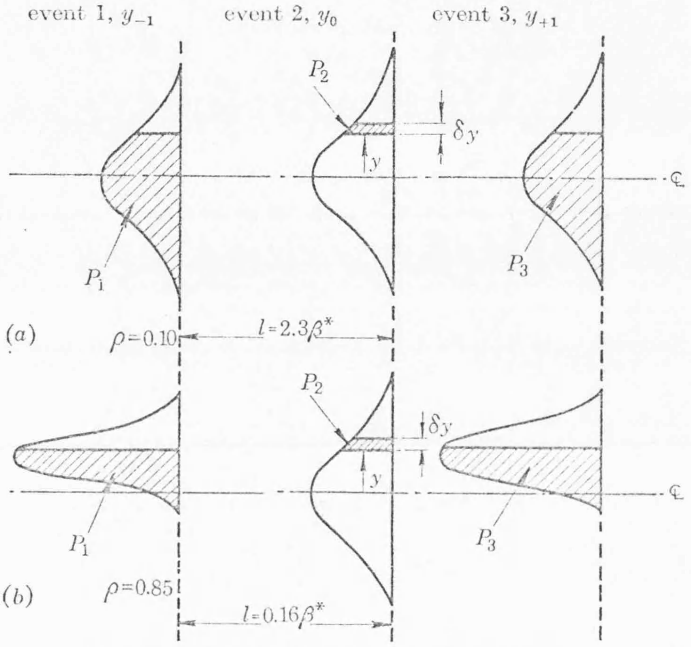


FIGURE 4. Model used in deducing distribution of peaks. (a) Sampling interval, $l = 2.3\beta^*$; correlation, $\rho = 0.10$; (b) sampling interval, $l = 0.16\beta^*$; correlation, $\rho = 0.86$.

Using this simple definition of a peak and equation (1) to define the probability density of a height distribution we can show (see the appendix) that the probability density of an ordinate being a peak at height y is given by

$$f^*(y) = [1/4\surd(2\pi)]\,[1+\mathrm{erf}\,(y/\surd 2)]^2 \exp\,(-\tfrac{1}{2}y^2),$$
$$= [1/\surd(2\pi)]\,\Phi^2(y) \exp\,(-\tfrac{1}{2}y^2), \qquad (5)$$

where here, and subsequently, f^* is used to indicate that properties of peaks are considered, f being retained for properties of the whole profile.

This argument can be extended to include the situation where the sampled ordinates are taken too close to be considered independent of each other. Such a situation is shown in figure 4b. In this case, also, the probability of an ordinate being a peak at height between y and $y+\delta y$ is given by $P_1P_2P_3$. However, the ordinates adjacent to the central y_0 are now not allowed to take all the values of the original height distribution; they take values that fit into a modified distribution whose shape depends upon the sampling interval. The modified height distributions shown in figure 4*b* are the result of having ordinates y_{+1} and y_{-1} so close to the central ordinate y_0, taken as a reference, that they are dominated by it. Thus, for short sampling intervals, y_{+1} cannot differ greatly from y_0 because of the inability of the signal level to change rapidly; this, in turn, is a consequence of the limitation of the power spectrum at high frequencies (figure 3*b*). In general, the distribution of y_{+1} is influenced not only by y_0, but also, to a lesser degree, by y_{-1}. However, for the particular example of the model used here (an exponential correlation function) y_{+1} can be considered as influenced by y_0 only and not by y_{-1}; this is a property indicative of a first order Markov process (Bendat 1958, p. 215) and the distributions of y_{+1} and y_{-1} drawn in figure 4*b* are dependent only upon the sampling interval and the value of y_0. (See the appendix, where the derivation of the equations of this section is outlined and the basis of a theory for any form of correlation function is indicated.)

On the more general theory associated with figure 4*b*, the probability density of an ordinate being a peak at height y is given by

$$\begin{aligned} f^*(y,\rho) &= \frac{1}{4\sqrt{(2\pi)}}\left[1+\operatorname{erf}\left(\frac{y}{\sqrt{2}}\sqrt{\frac{1-\rho}{1+\rho}}\right)\right]^2 \exp\left(-\tfrac{1}{2}y^2\right) \\ &= \frac{1}{\sqrt{(2\pi)}}\,\Phi^2\left(y\sqrt{\frac{1-\rho}{1+\rho}}\right)\exp\left(-\tfrac{1}{2}y^2\right). \end{aligned} \tag{6}$$

Figure 5 shows plots of this peak height distribution for a high and a low value of the correlation, ρ; the height distribution of the ordinates is also shown for comparison. The trends with varying values of ρ will be observed. As $\rho \to 0$ (large sampling interval), the shape of the peak height distribution becomes slightly skewed, its mean value approaches $+0.85$ and its standard deviation approaches a value of 0.70 Thus, when one uses larger sampling intervals the main, longer wavelength, structure of the profile is revealed (neglecting, here, the problem of aliasing (Bendat 1958)) and the peaks tend to lie above the centre line. As $\rho \to 1$ the shape of the peak height distribution, and its mean value and standard deviation, approaches those of the height distribution of the ordinates. Thus, when using short sampling intervals one is concerned with the shorter wavelength structure of the profile and therefore the peaks revealed by three point analysis follow closely the broad scale surface structure (cf. figure 1*b*, *c*).

The mean value of the peak height density curve $y^*(\rho)$ is found by taking the first moment of $f^*(y,\rho)$ in the normalized version of equation (6) and yields

$$\bar{y}^*(\rho) = \frac{1}{2N}\left(\frac{1-\rho}{\pi}\right)^{\frac{1}{2}}. \tag{7}$$

Similarly, the variance of the peak heights is the second central moment. Thus

$$[\sigma^*(\rho)]^2 = \left[1-\frac{(1+\rho)^{\frac{1}{2}}}{2N\pi\tan^2(N\pi)}-\frac{(1-\rho)}{4\pi N^2}\right], \tag{8}$$

where the normalizing factor N, giving the ratio of number of peaks to number of ordinates is

$$N = (1/\pi)\tan^{-1}\sqrt{\{(3-\rho)/(1+\rho)\}}. \qquad (9)$$

It will be noted that equations (5) and (6), when divided by equation (9), are the probability densities of peak heights. Equation (9) shows that as the correlation, ρ, increases from zero to unity, N falls from $\frac{1}{3}$ to $\frac{1}{4}$. These limiting values have a simple explanation. As the sampling interval is increased, $\rho \to 0$ and $N \to \frac{1}{3}$; the three events are then effectively independent (figure 4*a*) and the chance that any one of them (e.g. the centre one) is the highest becomes one-third. On the other hand, as the sampling interval is decreased $\rho \to 1$ and $N \to \frac{1}{4}$. The modified distributions of the two outer events are now centred upon the central ordinate (figure 4*b*); the areas P_1 and P_3 have values of $\frac{1}{2}$ and the probability that the central event is a peak is $\frac{1}{4}$.

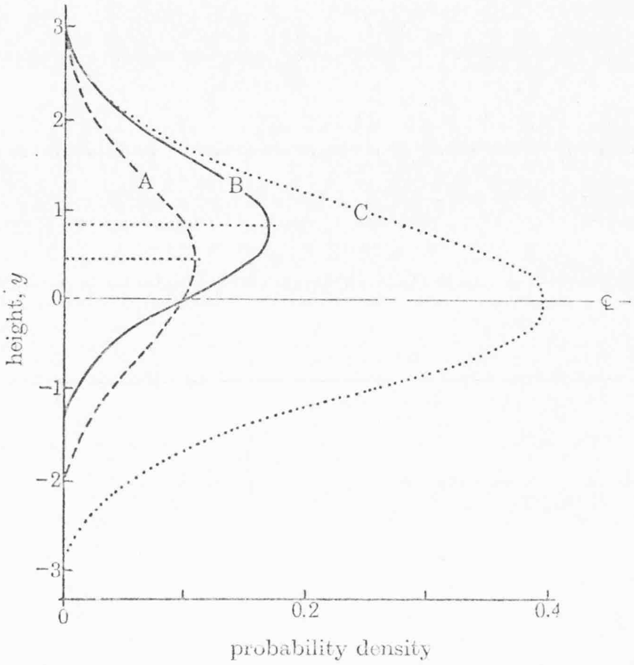


FIGURE 5. Probability densities of an ordinate being a peak at height y. The height, y, is normalized by the r.m.s. value (σ) of the ordinates. Results are shown for two different values of the sampling interval (l). (A) $l = 2.3\beta^*$; correlation, $\rho = 0.10$; average peak height $= 0.82$; (B) $l = 0.16\beta^*$; correlation, $\rho = 0{\cdot}86$; average peak height $= 0{\cdot}41$; (C) Gaussian distribution of ordinates.

To provide an adequate description of a surface in terms of a distribution of asperities it is also necessary to specify their radii of curvature. It is more convenient to discuss this in terms of a distribution of curvatures and we shall follow Greenwood & Williamson (1966) in deriving curvatures from the digital presentation of the profile. The assumption here is that one is justified in fitting a parabola to the profile by three point analysis; the problems involved in this assumption and the limits within which this analysis is justified will be discussed later. Consider, first, the example of three independent events (figure 4*a*). Figure 6*a* shows one possible arrangement of the three events which will give a peak at height y with a curvature C given by

$$C = 2y_0 - y_{+1} - y_{-1}. \tag{10}$$

In this equation C is non-dimensional but the true value of curvature depends upon the sampling interval l; to obtain a true value, C must be multiplied by σ/l^2. We designate the sampling interval as l, rather than β of equation (2), to emphasize the important point that the properties of the profile are being deduced by finite difference methods from data obtained by a sampling technique.

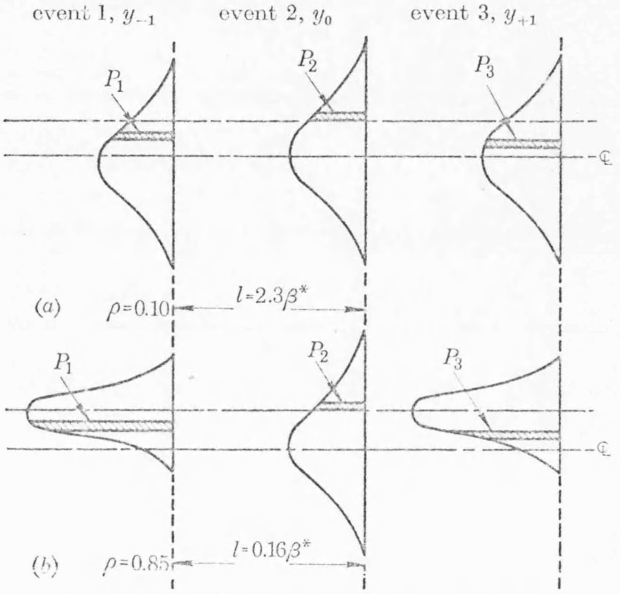


FIGURE 6. Model used in deducing the distribution of curvatures. (*a*) Sampling interval, $l = 2.3\beta^*$; correlation, $\rho = 0.10$; (*b*) sampling interval, $l = 0.16\beta^*$; correlation, $\rho = 0.86$.

This treatment assumes that the second derivative of the profile is an acceptable approximation to the curvature. Then the probability of the configuration shown in figure 6*a* is $P_1 P_2 P_3$, where P_1, P_2 and P_3, are given by the shaded areas shown. In order to obtain the *total* probability of a peak with curvature C at a height between y and $y + \mathrm{d}y$ many configurations, similar to that shown in figure 6*a*, must be taken into account. It is shown in the appendix that this total probability is given by a convolution integral. Thus the probability density function that any ordinate is a peak of curvature C at height y is

$$f^*(y, C) = \frac{\exp\left(-\frac{1}{2}y^2\right)}{2\pi\sqrt{2}} \exp\left[-\left(y - \tfrac{1}{2}C\right)^2\right] \operatorname{erf}\left(\tfrac{1}{2}C\right). \tag{11}$$

As before, the argument can be repeated for a shorter sampling interval (figure 6*b*). Thus

$$f^*(y, C, \rho) = \frac{\exp\left(-\frac{1}{2}y^2\right)}{2\pi[2(1-\rho^2)]^{\frac{1}{2}}} \exp\left[-\frac{[(1-\rho)\,y - \frac{1}{2}C]^2}{(1-\rho^2)}\right] \operatorname{erf}\left[\frac{C}{2\sqrt{(1-\rho^2)}}\right]. \tag{12}$$

The probability density function that any ordinate is a peak of curvature C (at any height) is obtained by integrating equation (12) to give

$$f^*(C,\rho) = \left[\frac{1}{4\pi(3-\rho)(1-\rho)}\right]^{\frac{1}{2}} \exp\left[\frac{-C^2}{4(3-\rho)(1-\rho)}\right] \operatorname{erf}\left[\frac{C}{2\sqrt{(1-\rho^2)}}\right]. \tag{13}$$

This distribution is skewed towards zero curvature; this is in general accord with the distribution of curvatures obtained by Greenwood & Williamson (1966) from digital analysis of a bead blasted surface. These authors suggest a Γ function as a suitable description of the distribution but equation (13) is nearer to a Rayleigh distribution than a Γ function and for large curvatures is very nearly Gaussian. A further comparison of these equations with results derived from surface profiles is given below.

The mean curvature $\bar{C}^*$ for all peaks is obtained by finding the first moment of $f^*(C,\rho)$ in equation (13). This yields

$$\bar{C}^* = (3-\rho)(1-\rho)^{\frac{1}{2}}/2N\sqrt{\pi}, \tag{14}$$

where the distribution has been normalized by N, the ratio of peaks to ordinates (equation (9)). The curvature (or, more strictly, the second differential) of the *profile* as a whole is given by

$$f(C,\rho) = \frac{1}{[4\pi(3-\rho)(1-\rho)]^{\frac{1}{2}}} \exp\left[\frac{-C^2}{4(3-\rho)(1-\rho)}\right], \tag{15}$$

which is equation (13) with the error function removed. This is a Gaussian distribution having a mean of zero and a standard deviation of $[2(3-\rho)(1-\rho)]^{\frac{1}{2}}$. A simple check on this value of the standard deviation (or, more strictly, the variance) is obtained by finding the square of the expected value of the curvature from equation (10). Thus $E[2y_0-(y_{-1}+y_{+1})]^2 = 6-8\rho+2\rho^2$.

The distribution of slopes is of importance because one widely used criterion for the onset of plastic flow (Blok 1952; Halliday 1955) uses the mean slope of the flanks of the asperities. The distribution of slopes (m) on the profile is easily obtained from the fact that the formula involves a simple linear relation of the Gaussian variates y_{-1} and y_{+1} and so is itself Gaussian with a mean of zero and a variance $|2\sigma^2(1-\rho^2)|/4l^2$. Hence

$$f(m,\rho) = \exp\left[\frac{-m^2l^2}{\sigma^2(1-\rho^2)}\right] \Big/ [4\pi(1-\rho^2)]^{\frac{1}{2}}, \tag{16}$$

from which one derives the average upward or downward slope (mean value of the modulus)

$$\bar{m} = \frac{\sigma}{l}\left[\frac{1-\rho^2}{\pi}\right]^{\frac{1}{2}}. \tag{17}$$

These formulae can also be obtained by the same type of procedure as that used in the derivation of equation (15).

4. Results of analysis of surface profiles

The validity of the theory given above has been checked by digital analysis of profile meter outputs. In order to present a coherent picture we give below a fairly complete analysis of the results obtained from one surface. The surface chosen (Aachen 64–13) is a typical ground surface used in an O.E.C.D. cooperative research programme (O.E.C.D, to be published). The surface profiles were taken at right angles to the direction of grinding.

The main experimental results were derived from surface profiles measured on a Talysurf 4 stylus surface roughness instrument in which a lightly loaded diamond stylus is traversed across the surface under examination. A normal stylus with a nominal tip dimension of 2.5μm was used. In these experiments the stylus was a diamond in the form of a truncated pyramid and the term 'tip dimension' refers to the linear dimension of the flat region at the tip in the direction of motion. A horizontal magnification of 500 was obtained by using the 500 × drive unit. Coupled to the Talysurf was a data logging system, specially devised for it, consisting of a Solartron A/D convertor and serializer together with a Data Dynamics 110 paper tape punch. From this equipment the amplified analogue signal of the surface profile was converted into a sequence of ordinates on the tape. Using this system ordinates were sampled at intervals of 1μm. The information on the paper tape was subsequently processed in an ICL 1905 computer.

Additional results were also obtained, with a special sharp stylus having a tip dimension of 0.25μm, the Talystep instrument being coupled to the same data logging equipment. The Talystep apparatus is capable of vertical magnifications of up to 10^6 and horizontal magnifications of up to 2×10^3. This latter facility made it possible to sample the profile at intervals of 0.25μm. Another feature of this instrument is the normal load on the stylus which was reduced to only 10^{-3} g (one-hundredth of the load on the Talysurf stylus) and this made the use of a very sharp stylus possible. Figure 7 shows an electron micrograph of this sharp stylus.†

A typical set of Talysurf results, from a single profile, consisted of some 10000 ordinates with a sampling interval of 1μm; of these, some 7000 were available for use after filtering to remove the d.c. level, the general slope and wavelengths comparable with the length of the profile. The results presented below are based upon five profiles and statistical analysis shows that the normalized standard errors are *ca.* 2% for mean values (e.g. equations (7) to (9)) and *ca.* 5% for points on the probability distributions (e.g. equations (4), (5) and (13)).

By suitable selection of data it was possible to present results for sampling intervals between 0.25 and 15μm. It was thus found (figure 8) that the model used in this paper was a good representation of the data obtained from the surface profiles; the distribution of ordinates was very close to Gaussian, with an r.m.s. value (σ) of 0.5μm, and the autocorrelation function was close to exponential

† We are indebted to Mr J. Jungles (Research Department Rank Precision Industries Ltd) for his skill in making and measuring this stylus.

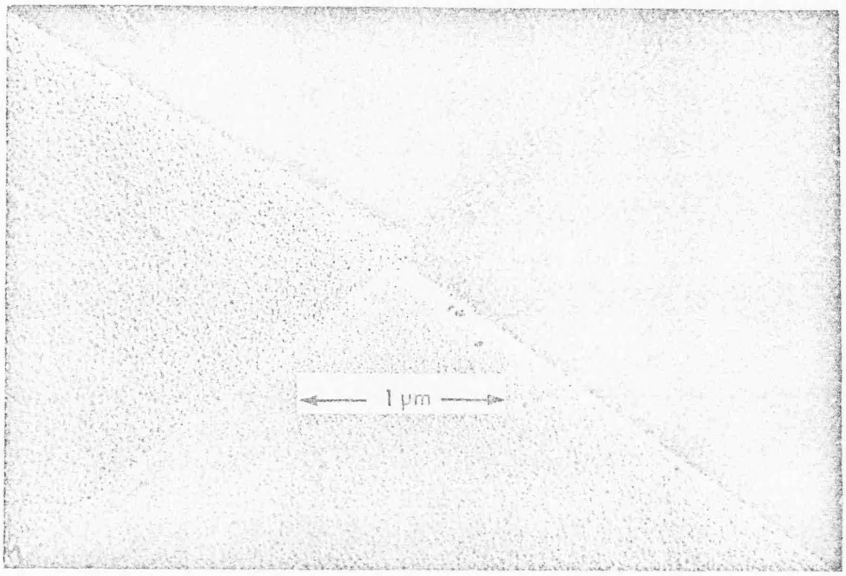


FIGURE 7. Electron micrograph of sharp stylus used in the experiments.

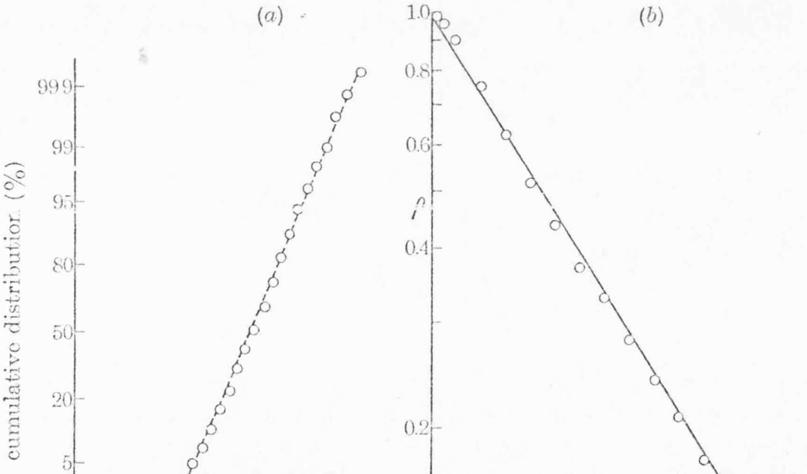


FIGURE 8. Characteristics of profiles of Aachen 64–13 represented as a random signal. (*a*) Cumulative distribution of heights (normal probability paper); (*b*) Correlation as a function of sampling interval (logarithmic-linear plot).

with a correlation distance (β^*) of 6.5μm. In subsequent discussion we shall use theoretical values of the correlation (ρ) for the selected values of the sampling interval as shown in table 1. It will be observed that any divergences between these values and those obtained from the profiles are, for the most part, within the limits of experimental error. In the results, presented below in graphical form, sampling intervals (l) of 15, 3 and 1μm (corresponding to correlations (ρ) of 0.10, 0.63 and 0.86 respectively) have been selected to display certain important features.

TABLE 1. RELATION BETWEEN SAMPLING INTERVAL (l) AND CORRELATION (ρ) BETWEEN SUCCESSIVE SAMPLES FOR AACHEN 64–13

sampling interval, l/μm	15	6.0	3.0	2.0	1.0	0.5	0.25
correlation, ρ	0.10	0.40	0.63	0.74	0.86	0.92	0.96

Figure 9 shows a comparison of theory and experiments for the probability that an ordinate is a peak at a height y (equation (6)). It will be observed that for $l = 15\mu$m and $l = 3\mu$m the agreement between theory and experiment is remarkably good. However, for $l = 1\mu$m (figure 9*c*) there is a marked divergence, the number of peaks detected falling significantly below the theoretical values. The results for all values of the sampling interval are shown in figure 10 in which the mean value and the standard deviation of the distribution of peaks (equations (7) and (8)) are plotted against the value of the correlation between successive samples. The most significant divergence between theory and experiment is the fact that, for the shorter sampling intervals, the mean values lie above the theoretical predictions (see also figure 9*c*).

Figure 11 presents theory and experiment for the probability than an ordinate is a peak of given curvature. As before, for $l = 15\mu$m and $l = 3\mu$m the agreement is excellent but there are significant differences for the shorter sampling interval of $l = 1\mu$m. It will be observed from the magnitudes of the curvatures shown in figures 11*a* to *c*, that, as the sampling interval is decreased, one is concerned with asperities of smaller and smaller radius. This is made quite clear in figure 12 which compares theoretical values of the mean curvature of the peaks with the values found from the profiles for differing sampling intervals. Once more, the only significant divergence between theory and experiment occurs at the shortest sampling interval ($l = 1\mu$m).

The results obtained at shorter sampling intervals (see, in particular, figures 9*c*, 11*c*) suggest that the measurements of the surface profiles are affected by the finite size of the stylus. In figure 11*c* a value of the nominal stylus curvature has been indicated; this is taken as the reciprocal of the nominal tip dimension of the stylus. The character of the divergence between theory and experiment shown in figure 11*c* is certainly consistent with the assumption that it arises from the finite size of the stylus. The total number of peaks detected is less than that forecast by the theory and the distribution has apparently been distorted towards smaller values of the curvature.

It is, of course, equally possible that the surface used in this work does not conform, at these shorter wavelengths, to the model assumed in this paper. Figures 9*c* and 11*c* would then imply that the structure of shorter wavelengths, although present in the model, does not exist upon the surface. In an attempt to resolve this question experiments were performed with a stylus with a smaller tip dimension. The results are shown in figure 13 where the ratio (N) of peaks to ordinates is plotted against the correlation (ρ) between successive samples. It will be recalled that the theory (equation 9) forecasts that this ratio varies between 0.33 ($\rho = 0$) and 0.25 ($\rho = 1$). Figure 13 shows, once more a divergence between theory and experiment for sampling intervals of less than $2\,\mu$m; in this region the

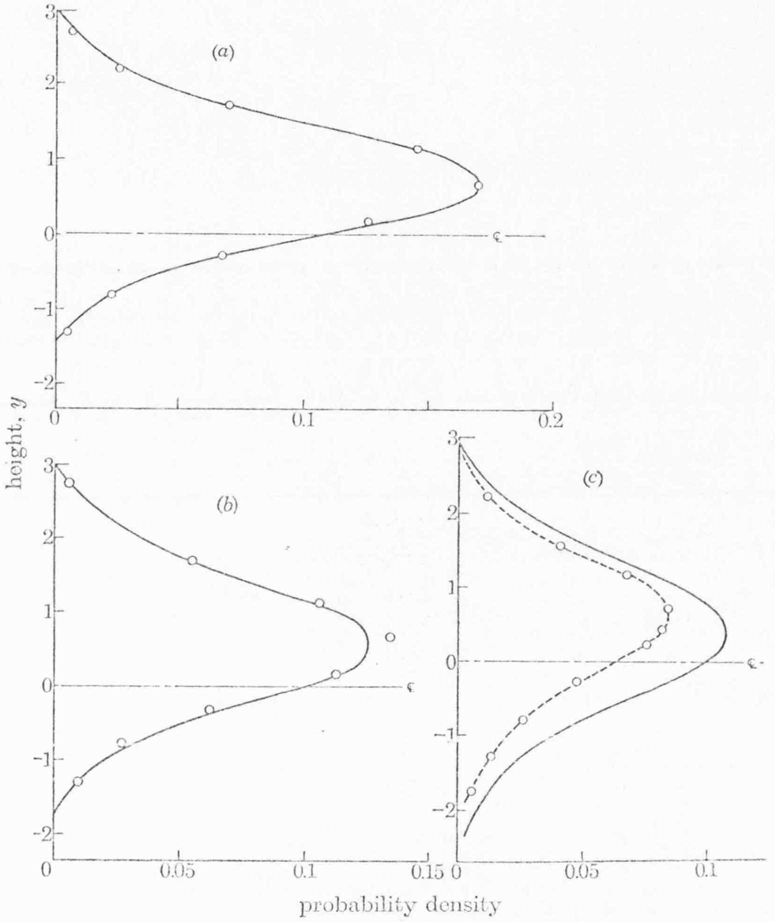


FIGURE 9. Probability densities of an ordinate being a peak at a height y. The full lines give the theory (equation (6)). The experimental points are derived from digital analysis of profiles of Aachen 64–13. Results are shown for the values of the sampling interval (l) corresponding to differing values of the correlation (ρ) between successive samples. (*a*) $l = 15\,\mu$m, $\rho = 0.10$; (*b*) $l = 3.0\,\mu$m, $\rho = 0.63$; (*c*) $l = 1.0\,\mu$m, $\rho = 0.86$.

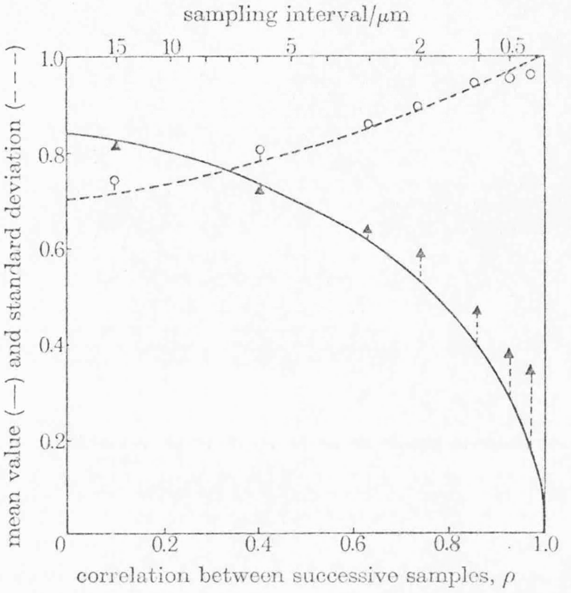


FIGURE 10. Characteristics of the distribution of peaks. The full line gives the mean value (equation (7)) and the broken line the standard deviation (equation (8)); they are normalized by the standard deviation of the ordinates (σ). The experimental points are derived from digital analysis of profiles from Aachen 64–13.

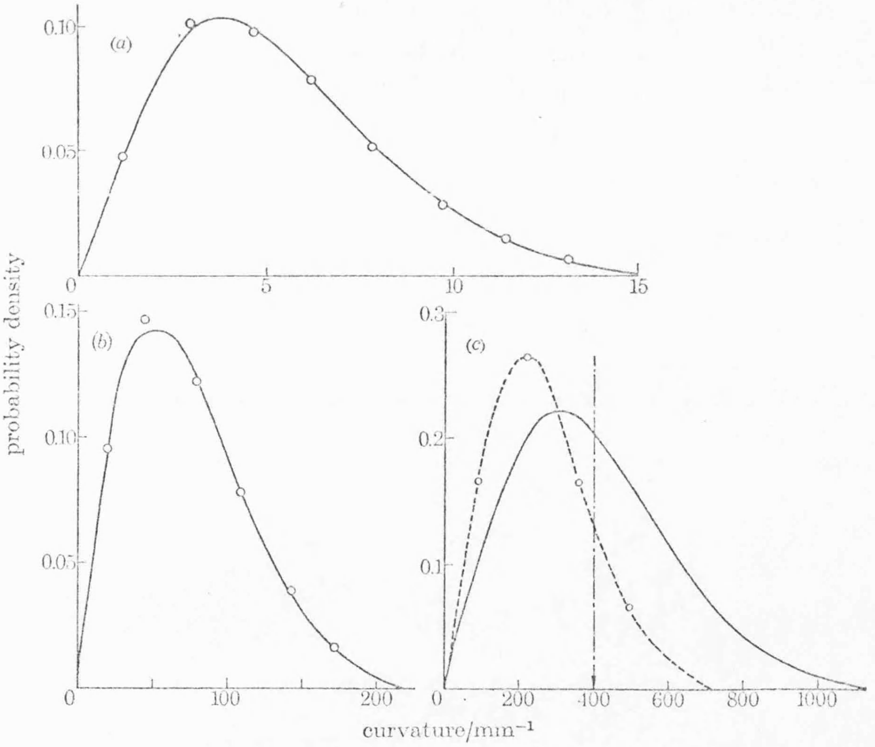


FIGURE 11. Probability densities of an ordinate being a peak of a given curvature. The full lines give the theory (equation (13)). The experimental points are from digital analysis of profiles from Aachen 64–13 (σ = 0.5 μm, β^* = 6.5 μm). Results are shown for three values of the sampling interval (l) corresponding to differing values of the correlation (ρ) between successive samples. (*a*) l = 15 μm, ρ = 0.10; (*b*) l = 3.0 μm, ρ = 0.63; (*c*) l = 1.0 μm, ρ = 0.86; the arrow indicates the nominal stylus curvature, 400 μm^{-1}.

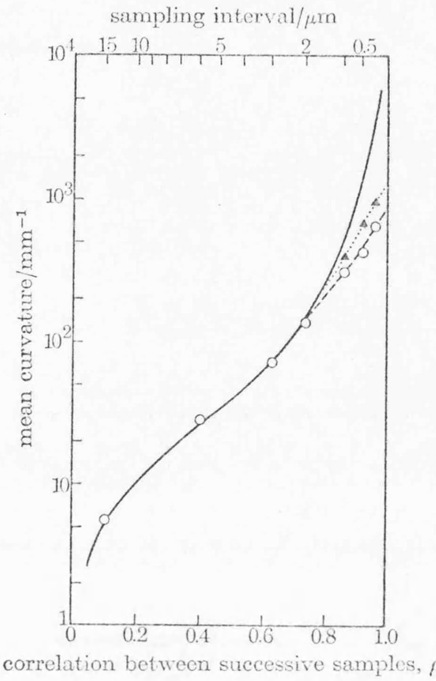


FIGURE 12. Mean curvature of the peaks as a function of the correlation (ρ) between successive samples. The full line gives the theory (equation (14)). The experimental points are derived from digital analysis of profiles from Aachen 64–13 ($\sigma = 0.5$ μm, $\beta^* = 6.5$ μm). ○, Normal stylus, nominal tip dimension 2.5 μm; ▲, special stylus, nominal tip dimension 0.25 μm.

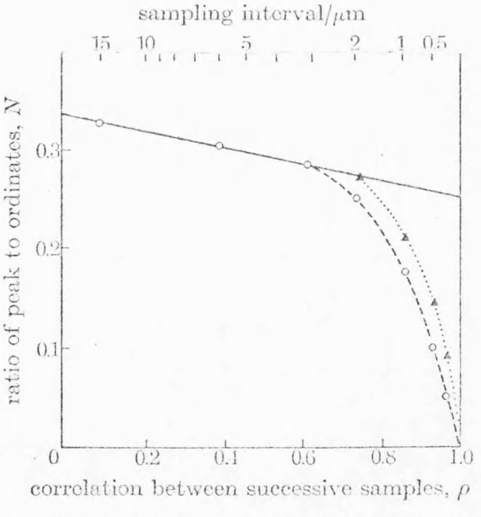


FIGURE 13. Ratio (N) of peaks to ordinates as a function of the correlation (ρ) between successive samples. The full line gives the theory (equation (9)). The experimental points are derived from digital analysis of profiles from Aachen 64–13. ○, Normal stylus, nominal tip dimension 2.5 μm; ▲, special stylus, nominal tip dimension 0.25 μm.

numbers of peaks detected falls well below the theoretical values. Similar plots showing a decline in the number of peaks detected at short sampling intervals have been presented by Sharman (1967–8), but no explanation of the cause has been advanced. Figure 13 also shows that when using a stylus with a smaller tip dimension the decline is delayed to smaller values of the sampling intervals. Clearly, therefore, stylus resolution is a significant factor affecting the behaviour in this region.

5. Discussion

The starting-point of the present work is the concept, well accepted in the theory of random processes but unexplored in the field of surfaces and their contact, that a random profile can be completely defined by two parameters: the height distribution and the autocorrelation function. For the particular example of the model used in this paper, these two parameters become simply two lengths, the r.m.s. value of the height distribution (σ) and the correlation distance (β^*). The statistical distributions of all significant geometric characteristics of the surface profile, for example, slopes, peaks and curvatures, can then be predicted from these two independent parameters. It is significant that the present work, which has arisen from markedly theoretical considerations, is set against a particular background of comment from those concerned with practical aspects of the measurement of surface finish and its use in engineering. This includes many comments that a measurement of surface finish, widely adopted hitherto, based solely upon the height distribution (r.m.s. or c.l.a. roughness) is not an adequate description of the functional significance of surface roughness (Reason 1967–8). Therefore, the ideas outlined here seem very relevant in any attempt to provide a more complete specification of surface finish. In practice, for reasons connected with ease of measurement, it may be desirable to measure a derived parameter such as the average upward or downward slope. However, although many surfaces do not conform exactly to the simple model used in this paper, the principle of specifying surface finish in terms of two independent parameters seems capable of wider application.

The importance of β^* in the specification of surface finish has been stressed, but it has an additional significance in its own right. Consider the measurement of the c.l.a. or r.m.s. roughness value of a random surface; it is desired to know the confidence limits of such a measurement and this is easily expressed if one knows the standard deviation of a large number of such measurements made upon the same surface. Alternatively this can be deduced if β^* is known. It can be shown that the standard deviation of such measurement of the c.l.a. roughness of a random surface, when normalized as a ratio of the mean value, is $\approx 1/\sqrt{(2M)}$. In this expression M is the ratio of the assessment length (L), used in the measurement of c.l.a., to the distance between points on the surface ($2.3\beta^*$) which just provides effectively independent events. (This ratio M is analogous to the band width × duration product used in communications theory to estimate the reliability of data.) For example, on the 0.03 in cut-off range, the Talysurf 4 instrument has an assessment length, of 0.15 in (3.81 mm). The value of $2.3\beta^*$ for Aachen 64–13 is 15μm. Thus the normalized standard deviation of c.l.a. readings on this surface should be 4.5 %. A measured value of the standard deviation for Aachen 64–13, based upon a large number of readings, was 4.3 %.

The comparison of theory and experiment outlined in this paper shows that the model which has been adopted can provide a satisfactory description of the geometric features of profiles from a surface typical of engineering practice; the statistical distribution of surface characteristics are accurately forecast over a wide range. This range covers an order of magnitude in the linear dimensions of the asperities and more than two orders of magnitude in their curvatures. Divergences appear only at shorter wavelengths and these arise, at least in part, from the resolution of the stylus. This is clear from a more detailed consideration of the results obtained with the sharp stylus.

First, consider the results of figure 13. The model adopted in the present paper requires that at short sampling intervals ($N \to \frac{1}{3}$) the number of peaks detected on a given length of profile should be inversely proportional to the sampling interval. However, with the normal stylus, reducing the sampling interval from 1 to 0.5μm increases the number of peaks by only 16 % and a further reduction to 0.25μm causes no detectable increase in the number of peaks. These results suggest that either the fine scale structure does not exist upon the surface, or it is present and is not detected by the stylus. The results obtained with the sharp stylus show, clearly, that stylus resolution is a significant factor. At a sampling interval of 1μm the replacement of the normal stylus by the sharp stylus causes an increase of 20 % in the number of peaks detected. In addition, when using the sharp stylus, a reduction in the sampling interval to 0.5μm and to 0.25μm causes increases in the number of peaks by 37 and 75 % respectively. Figure 12 shows that use of the sharp stylus also results in an increase in the mean curvature of the peaks; in other words the sharp stylus reveals more detail and finer detail. The suggestion that the finite size of the stylus has an influence upon the resolution of the finest detail is supported by other experiments (R. E. Reason, unpublished) in which a sharp stylus has revealed considerable detail upon smooth surfaces which the normal stylus does not reveal.

A detailed discussion of the effect of the size and shape of the stylus upon resolution is outside the scope of the present paper. However, some general considerations are worthy of comment. First, the effect of the finite size of the flat tip of the stylus is not likely to produce a sharp cut-off in resolution but should exercise an influence over a range of wavelengths somewhat as has been indicated in the discussion of figure 13. Nevertheless, assuming that the profile corresponds to the model of this paper, the change in the stylus tip dimension from 2.5 to 0.25μm produces a smaller change in the resolution than might be expected; perhaps the fine scale structure is present but its magnitude is less than is required by the theoretical model. However, it should be noted that the tip dimension may not be the only factor which determines the resolution of the stylus. Equation (17) shows that as the sampling interval is reduced, and structures of shorter wavelength are involved, the local slope of the surfaces becomes steeper (cf. figures 1 *a*, *c*). Thus, although existing instruments reveal the details known to be of functional significance, to resolve the finest detail on random surfaces it may be necessary to consider the complete shape of the stylus, the sides as well as the tip.

The present work has also focused attention upon a problem associated with the representation of random surfaces by models. It has been explained that the random signal representation used in this paper is a complete description of the profile, in a statistical rather than a deterministic sense. However, models appropriate to the problems of surface contact are expressed as a distribution of peaks and such models have been derived from digital analysis of surface profiles; consideration of the results obtained in this paper shows, immediately, the difficulties associated with any definition of a peak. The simpler forms of digital analysis (for example, the three-point system which has been widely used) cause, inevitably a loss of information compared with that provided in the original random signal representation, or that which is present in the surface profile. Thus, a close sampling technique collects the maximum amount of information about the profile; but three-point analysis of these data restricts the information to structures of this same small order of size. Only by the use of more sophisticated techniques, such as digital filtering, can the total information in the sample be utilized. An alternative is to use differing selections of the same information, by rejecting some data, but still to use three-point analysis; e.g. to present, as we have done, information from three-point analysis for a wide range of sampling intervals.

However, these techniques present problems of rigour. When using the longer sampling intervals one is presenting statistical information about the longer wavelength structure of the profile obtained by drawing a smooth curve through widely separated points upon it. Therefore the results presented above, in particular, the values of curvature obtained from long sampling intervals, should be interpreted with these reservations in mind.

Because the description of the profile as a random signal is statistical in its form an overemphasis upon the limitations of three-point sampling is not, perhaps, the most significant problem. The more important, and difficult, question is deciding the range of sampling intervals from which it is acceptable to use information in devising models for theories of surface contact. The significance of the shorter wavelength structures (revealed by the use of shorter sampling intervals) in the contact and rubbing of surfaces may be questioned; in any event, a lower limit to the acceptable range of sampling intervals is set by stylus resolution. At the other extreme it is clear that the use of very long sampling intervals will give results which have little physical significance.

A more relevant question is the spacing of events that will define the dominant or main structure of the profile; such a spacing might be considered as the upper limit of sampling intervals to be used in devising models for the study of surface contact. To outline this structure implies that the profile be considered as a series of events which have just reached the spacing where they can be regarded as effectively independent. As discussed earlier, this suggests a spacing of events of $2.3\beta^*$ and also implies that the main structure of the profile has a wavelength of *ca.* $10\beta^*$. Similarly the profile can be considered in terms of the power spectral density function (figure 3*b*). This suggests that the most significant information is contained in a low-pass band of frequencies having a cut-off frequency of $(2\pi\beta^*)^{-1}$. This approach suggests a wavelength of *ca.* $6\beta^*$ for the main structure of the profile. Finally, we can arrive at an estimate of the broad scale structure from a consideration of the number of times the profile crosses the centre line; the mean distance between such crossings might be regarded as half the wavelength of this main structure. To obtain a meaningful result with this approach one must remove the high frequency components. If we replace the power spectral density distribution of figure 3*b* by one with a sharp cut-off at an angular frequency of $1/\beta^*$ (low-pass filtered white noise) it is possible to use a result derived by Rice (see Bendat 1958, p. 128). For this situation the mean distance between centre line crossings is $5.5\beta^*$ and a representative value of the wavelength of the main structure is $11\beta^*$. It can be argued that, when considering the results of three-point analysis, a somewhat shorter interval is appropriate because in surface contact one is concerned with the tips of the higher asperities. As a general guide to the main structure of relevance in contact problems we shall therefore use the results derived from a sampling interval of $2.3\beta^*$.

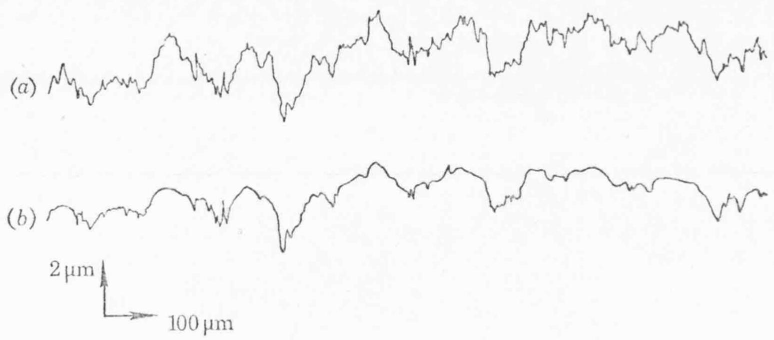


FIGURE 14. Talysurf profiles of cylindrical specimens used in lubricated friction experiments. (*a*) Original surface profile; (*b*) profile of the same line after one traversal of the load. Specimens 0.5 % C steel; 300 d.p.n.; 0.635 mm diameter; load 25 N.

The significance of this distinction between the differing scales of size involved in surface structure and the importance of the main, broad scale, structure is illustrated in figure 14. The upper record shows the original profile of a cylindrical specimen used in a low speed boundary lubricated friction experiment using a crossed-cylinders friction machine (Archard 1958). The lower profile is of the *same* *track* upon the surface after just one traversal of the load. It will be seen that the smaller scale structure has been lost by plastic flow during the first traversal but the longer wavelength structure remains upon the surface and is the dominant factor in determining the contact conditions in subsequent traversals of the surface. A more complete account of these experiments will be published elsewhere (D. J. Whitehouse & J. F. Archard, to be published).

An important aim of recent studies of the topography of surfaces has been to provide an estimate of the chances that a given surface will be subjected to plastic flow during contact. Blok (1952) and Halliday (1955) considered the shape of asperities which could be pressed flat without recourse to plastic deformation. It was shown that this criterion could be expressed in the form

$$\overline{m} \leqslant KH/E', \tag{18}$$

Where H is the hardness and $E' = E/(1-\nu^2)$, E being the Young modulus and ν the Poisson ratio; $\overline{m}$, as above, is the average upward or downward slope and K is a numerical factor, in the range 0.8 to 1.7, which depends only upon the assumed shape of the asperity. Greenwood & Williamson (1966) assessed the probability of plastic deformation using their model in which asperities, each of radius R, are disposed in a Gaussian distribution of heights of standard deviation σ^*. In this model there is always a finite chance of plastic flow; however, it was shown that it depended very little upon the load but was critically dependent upon a plasticity index, ψ, given by

$$\psi = \left(\frac{E'}{H}\right)\left(\frac{\sigma^*}{R}\right)^{\frac{1}{2}}. \tag{19}$$

The plasticity criteria of equations (18) and (19) are similar in form. The Blok–Halliday criterion (equation (18)) is, however, unduly severe because it assumes complete depression of the asperities. The plasticity index of Greenwood & Williamson (1966) takes account of the fact that only the tips of asperities are normally involved in contact. The present work emphasizes the simplifications which have been made in these plasticity criteria because they take no account of the existence, upon surfaces, of superposed asperities of differing scales of size. The plasticity calculations of equation (18) and (19) assume that the deformation of each of the asperities is independent. Therefore the plasticity index of equation (19) has a significance only if it is applied to the main, long wavelength structure; it should then indicate the probability of plastic flow over regions associated with this scale of size. If values of R corresponding to smaller scale structure are used the arguments involved in the derivation of equation (19) become invalid because the deformation of adjacent asperities interact, as in the model of figures 1*b*, *c*.

To summarize the results derived from the model of the present paper, table 2 shows the way in which the significant characteristics of a surface profile depend upon the two independent parameters σ and β^*. To emphasize the importance of the scale of size used in the analysis each characteristic (except ψ, for reasons outlined above) is shown for two scales. The main structure of the profile is derived assuming a sampling interval, l, of $2.3\beta^*$ and the fine scale structure assumes asperity dimensions one order of magnitude smaller ($l = 0.23\beta^*$). For Aachen 64–13, used in our digital analysis described above, this is just within the limits of resolution of the Talysurf instrument using the normal stylus. In deriving a value of the plasticity index, ψ, a value of the mean curvature of the peaks, derived from equation (14), has been used. The value of ψ derived in this way somewhat underestimates the probability of plastic flow because, as equation (11) shows, the curvature of the peaks increases with increasing height. Thus the highest peaks, which are those involved in contact, have a smaller radius than the total population.

TABLE 2. CHARACTERISTICS OF A RANDOM PROFILE IN TERMS OF σ AND β^*

characteristic of the profile	main structure ($l = 2.3\beta^*$)	fine structure ($l = 0.23\beta^*$)
mean of peak distribution	$+0.82\sigma$	$+0.47\sigma$
standard deviation of peak distribution, σ^*	0.71σ	0.9σ
ratio of peaks to ordinates, N	0.33	0.26
average upward or downward slope, $\overline{m}$	$0.24\sigma/\beta^*$	$1.66\sigma/\beta^*$
mean peak curvature, $\bar{c}^*$	$0.45\sigma/\beta^{*2}$	$20\sigma/\beta^{*2}$
plasticity index, ψ	$0.3\left(\frac{E'}{H}\right)\left(\frac{\sigma}{\beta^*}\right)$	—

A complete theory of the contact of surfaces on the basis the model of this paper cannot be presented here. However, one interesting feature of the model and its comparison with that of Greenwood & Williamson (1966) is worthy of comment. The Greenwood & Williamson model is specified by three parameters: σ^*, the standard deviation of the peak height distribution; R, the radius of curvature of the asperities; and η, the density of asperities per unit area. In the model of this paper the required parameters are completely defined by σ, the standard deviation of the height distribution and β^*, the correlation distance. The theory of contact based on our model involves a statistical distribution of both peak heights and peak curvatures. Comparing the two models: σ^* is proportional to σ, R is proportional to β^{*2}/σ, and η is proportional to $1/\beta^{*2}$. Therefore for all random surfaces, when the Greenwood & Williamson model is used the parameters should be related by the equation

$$\sigma^* R\eta = \text{constant}.$$

There is some evidence (J. A. Greenwood, private communication) from the analysis of bead-blasted surfaces that this relation is, indeed, obeyed.

Finally, a brief comment upon the generation, by mechanical processes, of surfaces having a random structure. Such surfaces are generated by multiple contacts between particles and the surfaces. Thus in grinding and sand blasting the unit event is the interaction of a grit with the surface resulting in the displacement or removal of material. If one postulates a random element in these events it seems likely that the surfaces thus generated would have a random structure in which the standard deviation of the height distribution bears a simple relationship to the depth of the unit event and the correlation distance similarly bears a simple relation to the width of the unit event.

We are indebted to Dr J. A. Greenwood and Dr J. B. P. Williamson for discussions of their published and unpublished work. Our thanks are due to our colleague Dr P. H. Phillipson for helpful discussions of the theory of random processes. We also thank the Directors of Rank Precision Industries Ltd for financial support and for permission to publish.

Appendix. General form of the theory

In the derivation of the relevant expressions use will be made of the multidimensional normal distribution (m.n.d.) which is concerned with the joint probability density function of a number of Gaussian variates; in our problem these variates will be simply profile ordinates. The m.n.d. will be Gaussian because any linear combination of Gaussian variates is itself Gaussian.

Definition (see Bendat 1958). If the ordinates $y_1, y_2, \dots, y_N$ have Gaussian height distributions, having a mean of zero and a variance unity then their combined joint density function is given by

$$f(y_1, y_2, \dots, y_N) = \frac{1}{(2\pi)^{\frac{1}{2}N}|M|^{\frac{1}{2}}} \exp\left[\frac{-\sum\limits_{i,j=1}^{N} M_{ij}y_iy_j}{2|M|}\right]. \tag{A1}$$

Where $|M|$ is the determinant of M; M is given by the square matrix

$$M = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ \dots & d_{ij} & \dots & \dots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{pmatrix},$$

d_{ij} being the second moment of the variables y_iy_j. M_{ij} is the co-factor of d_{ij} in M.

Take, for example, the joint probability density of two ordinates, say y_{-1} and y_0 correlated by ρ. Then

$$f(y_0, y_{-1}) = \frac{1}{\surd(2\pi)} \exp\left(-\tfrac{1}{2}y_0^2\right) \frac{1}{\surd\{2\pi(1-\rho^2)\}} \exp\left[-\frac{(y_{-1}-\rho y_0)^2}{2(1-\rho^2)}\right].$$

Similarly, for three ordinates, y_{-1}, y_0, y_{+1}, having a correlation of ρ_1 between adjacent ordinates and ρ_2 between extreme ordinates the joint probability density is given by

$$f(y_{-1}, y_0, y_{+1}) = f(y_0)\, f(y_{-1}/y_0)\, f[y_{+1}/(y_0, y_{-1})], \tag{A2}$$

where
$$f(y_0) = \frac{1}{\sqrt{(2\pi)}}\exp\left(-\tfrac{1}{2}y_0^2\right)$$

$$f(y_{-1}/y_0) = \frac{1}{\sqrt{\{2\pi(1-\rho_1^2)\}}}\exp\left[-\frac{(y_{-1}-\rho_1 y_0)^2}{2(1-\rho_1^2)}\right]$$

$$f[y_{+1}/(y_0, y_{-1})] = \frac{(1-\rho_1^2)}{\sqrt{\{2\pi(1-\rho_2)\,(1+\rho_2-2\rho_1^2)\}}}$$
$$\times\exp\left[-\frac{\{y_1(1-\rho_1^2)-y_0\rho_1(1-\rho_2)-y_{-1}(\rho_1^2-\rho_2)\}^2}{2(1-\rho_1^2)\,(1-\rho_2)\,(1+\rho_2-2\rho_1^2)}\right].$$

For the model under consideration the autocorrelation is exponential. Hence $\rho_1^2 = \rho_2$. Then $f[y_{+1}/y_0, y_{-1}]$ reduces to $f(y_{+1}/y_0)$ which is a criterion for a first-order Markov process. Hence for an exponential correlation function equation (A 2) becomes

$$f(y_{-1}, y_0, y_{+1}) = \frac{1}{\sqrt{(2\pi)}}\exp\left(-\tfrac{1}{2}y_0^2\right)\frac{1}{\sqrt{\{2\pi(1-\rho^2)\}}}$$
$$\times\exp\left[-\frac{(y_{-1}-\rho y_0)}{2(1-\rho^2)}\right]\frac{1}{\sqrt{\{2\pi(1-\rho^2)\}}}\exp\left[-\frac{(y_{+1}-\rho y_0)^2}{2(1-\rho^2)}\right]. \quad \text{(A 3)}$$

It will be noted that the simplified form of the theory used in this paper arises because $\rho_2 = \rho_1^2$ and for a *particular* sampling interval this result does not depend upon the fact that the autocorrelation function is exponential. However, the exponential correlation function is an essential requirement for the general applicability of the simplified theory at *all* sampling intervals.

Peak height distribution

The probability of an ordinate being a peak at a height between y and $y+dy$ is written in terms of a restriction of the joint probability density range in the following form (see theory):

$$\text{prob}\,[y_{-1} < y, y < y_0 < y+dy, y_{+1} < y]$$
$$= \int_{-\infty}^{y}\int_{y}^{y+dy}\int_{-\infty}^{y} f(y_{-1}, y_0, y_{+1})\,dy_{-1}\,dy_0\,dy_{+1}.$$

Which is the general equation of a peak using the three ordinate model; when the correlation is exponential the probability density reduces to the following form

$$f^*(y, \rho) = \frac{\exp\left(-\tfrac{1}{2}y^2\right)}{\sqrt{(2\pi)}}\,\frac{1}{2\pi(1-\rho^2)}$$
$$\times\int_{-\infty}^{y}\exp\left[\frac{-(y_{-1}-\rho y_0)^2}{2(1-\rho^2)}\right]dy_{-1}\int_{-\infty}^{y}\exp\left[\frac{-(y_{+1}-\rho y_0)^2}{2(1-\rho^2)}\right]dy_{+1}.$$

This reduces to equation (6) of the theory and, for $\rho = 0$, gives equation (5).

The ratio (N) of the number of peaks to ordinates, for any value of the correlation ρ, is obtained simply by integrating equation (6).

Peak curvature distribution

For a given curvature, C, as defined in the text, the ordinates are related by an expression

$$C = 2y_0 - y_{-1} - y_{+1}.$$

Hence the total probability of an ordinate being a peak between y and $y + \mathrm{d}y$ and describing a curvature, C covering all possible configurations of y_{-1} and y_{+1}, is given by

$$f^*(y, C, \rho) = \int_y^{y+\mathrm{d}y} \int_{y-C}^{y} f(y_0)\, f(y_{-1}/y_0)\, f\left[\frac{2y_0 - y_{-1} - C}{y_0, y_{-1}}\right] \mathrm{d}y_{-1}\,\mathrm{d}y_0. \qquad \text{(A 4)}$$

Hence the probability density of an ordinate being a peak at height y with curvature C is

$$f^*(y, C, \rho) = f(y) \int_{y-C}^{y} f[y_{-1}/y]\, f\left[\frac{2y - y_{-1} - C}{y, y_{-1}}\right] \mathrm{d}y_{-1}. \qquad \text{(A 5)}$$

This is a convolution integral which enables simple graphical equivalents to be constructed; equation (A 5) corresponds to the general curvature formula with a general autocorrelation function. However, for the simple exponential correlation it becomes

$$f^*(y, C, \rho) = \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{y^2}{2}\right) \int_{y-C}^{y} \frac{1}{2\pi(1-\rho^2)}$$
$$\times \exp\left[-\frac{(y_{-1} - \rho y)^2}{(12 - \rho^2)}\right] \exp\left[-\frac{(2y - y_{-1} - C - \rho y)^2}{2(1-\rho^2)}\right] \mathrm{d}y_{-1}.$$

This reduces to equation (12) of the theory and when $\rho = 0$ one obtains equation (11).

References

Archard, J. F. 1957 *Proc. Roy. Soc. Lond.* A **243**, 190.

Archard, J. F. 1958 *Wear* **2**, 21.

Beckmann, P. & Spizzichino, A. 1963 *The Scattering of electromagnetic radiation from rough surfaces*. London: Pergamon.

Bendat, J. S. 1958 *Principles and applications of random noise theory*. New York: Wiley.

Blok, H. 1952 *Proc. Roy. Soc. Lond.* A **212**, 480.

Bowden, F. P. & Tabor, D. 1954 *Friction and lubrication of solids*. Oxford University Press.

Greenwood, J. A. & Williamson, J. B. P. 1966 *Proc. Roy. Soc. Lond.* A **295**, 300.

Halliday, J. S. 1955 *Proc. Instn mech. Engrs* **109**, 777.

O. E. C. D. Sub-group Typology and Topology (to be published).

Peklenik, J. 1967–8. *Proc. Instn mech. Engrs* **182**, part 3K (conference on the properties and metrology of surfaces), p. 108.

Reason, R. E. 1967–8 *Proc. Instn mech. Engrs* **182**, part 3K (conference on the properties and metrology of surfaces), p. 300.

Sharman, H. B. 1967–8 *Proc. Instn mech. Engrs* **182**, part 3K (conference on the properties and metrology of surfaces), p. 416.

Presented to the A.S.M.E. Conference on
Surface Mechanics, Los Angeles, November 1969.

The Properties of Random Surfaces in Contact

D. J. Whitehouse

and

J. F. Archard

Abstract

Theories of the contact of rough surfaces use models in which the surfaces are represented as a distribution of asperities. Realistic models of surfaces should be based upon knowledge of the topography of surfaces used in engineering practice. In recent years much knowledge of surfaces has been gained from the digital presentation and analysis of the output of instruments in which the surface profile is explored by means of a lightly loaded stylus.

This paper explores the concept of a random signal as a model of the surface profile; such a model, expressed as a height distribution and an auto-correlation function, is a complete representation of the profile. It is shown that, within the limits where stylus resolution is not a significant factor, such a model is an adequate representation of the profiles of a typical ground surface. For surface contact studies it is then required to transform the random signal model into a form in which the surface is represented by a distribution of asperities. Suitable asperity models are discussed in the light of the theoretical analysis and the results of digital analysis of surface profiles. The significance of this work is illustrated by some results showing changes in surface finish caused by running in under conditions of boundary lubrication.

Introduction

Two concepts form the background of this paper. The first is the idea that the way in which surfaces touch is of vital importance in many branches of engineering; friction and wear, the conduction of heat and electricity between bodies in contact are examples of important areas of engineering practice where the nature of the true area of contact is of vital significance. The second concept is that all surfaces are rough; certainly all surfaces used in engineering practice consist of hills and valleys which are very large compared with atomic dimensions. The major problem in this field is to take the methods by which the roughness of surfaces are measured and characterized and to transform this information into a form in which it can be used in theories of surface contact.

Theories of Surface Contact

It is well accepted that, because surfaces are rough, the true area of contact is very much smaller than the apparent area in contact. Much of the development, over the post-war era, of a scientific approach to the study of friction, lubrication and wear [1] is based upon the idea that the pressures at true areas of contact are very large indeed. It is commonly assumed that the hills, or asperities, upon surfaces can be regarded as spherical caps. If the load on a single asperity is sufficiently small the deformation is elastic and the relationship between area of contact, A, and load, W, is given by the Hertz equations (2)

$$A = \pi \left[\frac{3WR}{2E'} \right]^{\frac{2}{3}} \tag{1}$$

where $E' = E/(1 - \nu^2)$; E is Young's modulus and ν is Poisson's ratio. R is the relative radius of curvature of the surfaces in contact. At heavier loads the deformation is plastic and the area of contact is given by [3]

$$A = W/H \tag{2}$$

where H is the hardness of the material.

In their earlier work Bowden and Tabor [1] suggested that the deformation of the asperities was plastic. Therefore the true area of contact was proportional to the load and, assuming that friction arises from the force required to shear these areas, the frictional force is also proportional to the load. Thus it was possible to provide a simple and elegant explanation of Amontons' laws of friction.

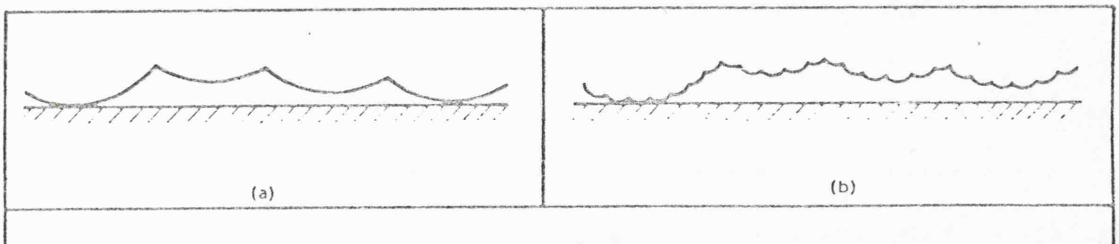


FIG. 1 MODELS OF SURFACES CONTAINING ASPERITIES OF DIFFERING SCALES OF SIZE. When the deformation is elastic the relationships between the area of contact (A) and the load (W) are as follows: (a) $A \propto W^{4/5}$, (b) $A \propto W^{14/15}$, (c) $A \propto W^{44/45}$.

In recent years it has been recognized that the contact of surfaces must involve an appreciable proportion of asperity contacts at which the deformation is entirely elastic. Archard [4,5] suggested a series of models (Figure 1) to represent surfaces. It was shown that as these models became more complex the relationship between *A* and *W*, assuming entirely elastic deformation, moved more closely towards direct proportionality. Thus it was shown that a satisfactory explanation of Amontons' laws of friction is not dependent upon the assumption of plastic deformation. Although it was admitted that these models were artificial, two features are worthy of comment. First, the geometric features of the particular model used may not be of the greatest significance in seeking an explanation of Amontons' laws. What is important is the physical consequences; in the complex models an increase in the load is almost entirely used in creating new areas rather than enlarging existing ones. Secondly, the models contain a feature to which we shall revert later; they assume that there exists upon the surface superposed asperities of widely differing scales of size.

In the most important recent contribution to the subject Greenwood and Williamson [6,7] used the results of an analysis of measured surface profiles. They showed that for many surfaces the asperities have a Gaussian distribution of heights. Their model of surfaces therefore consisted of a series of spherical asperities, each having the same radius of curvature as in the model of Fig. 1a, but having a Gaussian distribution of heights. With such a model the relationship between *A* and *W* is again close to direct proportionality. Thus a second theoretical derivation of Amontons' laws of friction was provided for elastic deformation under multiple contact conditions.

If, as in the earlier work of Bowden and Tabor [3], it is assumed that the deformation is entirely plastic, the details of the surface finish seem relatively unimportant since the total area of contact and the contact pressures do not depend upon surface topography. However, if an appreciable proportion of the load is borne by elastic deformation the role of surface topography becomes much more significant. For example, the proportion of the load which is borne by plastic flow, even though it be very small, may be highly significant in the initiation of various forms of surface failure.

The Measurement and Characterization of Surface Roughness

Many different methods have been used in the measurement and characterization of surface topography. These include optical, capacitance, and electron optical methods. All of these have a role in the examination of surfaces. However in this paper we shall confine ourselves to the most commonly used method in which a lightly loaded stylus is moved over the surface. Its vertical movement is converted into an electrical signal which is amplified. The output signal is displayed on a chart recorder, thus presenting a representation of the surface profile, or is used, after suitable filtering to provide a meter reading. In engineering practice an acceptable characterization of the surface roughness is often taken as the center line average (*cla*) or *rms* value of this signal. A typical profile together with the coordinate system used in this paper is shown in Fig. 2. In recent years it has been widely recognized, by those concerned with the measurement of surface finish and its use in engineering practice, that a characterization based solely upon *rms* or *cla* readings is sometimes not an adequate description of the functional significance of surface roughness [8].

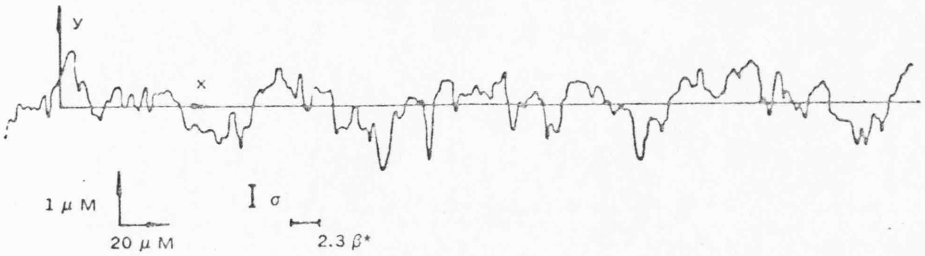


FIG. 2 SURFACE PROFILE OF AACHEN 64-13 SHOWING COORDINATE SYSTEM. The magnitude of the rms value of the height distribution (σ) and the correlation distance β* are shown for comparison with the profile.

One feature of stylus instruments which is of particular importance in the context of our present discussion is the ease with which the output can be analyzed by analogue and digital techniques. In the field of production engineering this information has been presented in many different ways [9] ; height distributions, slope distributions, power spectral density curves and autocorrelation functions are but a few of the characteristics which have been displayed. However, the relationship between these characteristics and the functional performance of surfaces has received little attention.

Greenwood and Williamson [6] , and many others, have shown that for many surfaces the distribution of heights is very close to Gaussian. Using digital analysis, Greenwood and Williamson also investigated the distribution of peak heights. The most common technique used in this type of investigation is three point analysis in which a peak is defined when the central of three successive sampled heights lies above those on either side. It was shown that the distribution of peaks was also close to Gaussian but the mean value and the standard deviation of this distribution differed from that of the heights of the ordinates. We shall consider later the significance of these findings.

A Random Signal Model of a Surface Profile

Since many surfaces have a distribution of heights which is Gaussian it seems appropriate to use the powerful techniques which have been used in the analysis of random processes as an aid in analysis of surface topography. We therefore consider as a model of a surface profile a random signal defined completely (in a statistical rather than a deterministic sense) by two parameters: a height distribution and an autocorrelation function. We shall confine our attention to the particular example of a surface profile having a Gaussian distribution of heights and an exponential correlation function. The major reason for using this model is that many surfaces used in engineering practice conform to this model or are a close approximation to it [10] . Moreover this model has been widely used in both the theory of random processes [11] and the representation of surfaces in studies of the scattering of electromagnetic radiation [12] .

Fig. 2 shows a typical profile and the coordinate system adopted here; the mean line through the profile will be taken as $y = 0$. In practice, the *dc* level, the general slope and curvature of the surface are removed by a filter removing the longest wavelengths. This does not substantially affect the autocorrelation function. If the distribution of heights is Gaussian the probability of finding an ordinate at a height between y

and $(y + dy)$ is $f_y \, dy$ where the height probability density function f_y is given by

$$f_y = \frac{1}{\sqrt{2\pi}} \exp \left(- \tfrac{1}{2} y^2\right) \qquad (3)$$

In this equation the x axis $(y = 0)$ is taken as the mean line through the profile and the ordinates have been normalized by the standard deviation, σ, of the height distribution.

The autocorrelation function is defined as

$$C(\beta) = \lim_{L \to \infty} \frac{1}{L} \int_{-L/2}^{+L/2} y(x) \cdot y(x+\beta) \, dx \qquad (4)$$

In this equation $y(x)$ is the height of the profile at a given coordinate and $y(x+\beta)$ is its height at an adjacent coordinate $(x+\beta)$. $C(\beta)$ is an expression of the way in which the statistical dependence of the heights of two adjacent points on the profile depends upon their spacing. We assume that

$$C(\beta) = \exp (\beta/\beta^*) \qquad (5)$$

where β^* will be called the correlation distance. Thus the statistical dependence of the heights of two points on the profile declines towards zero as their spacing, β, increases.

There is a well known Fourier transform relationship between the autocorrelation function, $C(\beta)$, and the power spectral density function of the waveform, $P(\omega)$, of a random waveform given by the equation

$$C(\beta) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} P(\omega) \exp (j\omega\beta) \, d\omega \qquad (6)$$

This relationship between $C(\beta)$ and $P(\omega)$ for the particular model used here is shown in Figure 3. The spectrum consists of white noise limited only at higher frequencies by a cut-off of 6 db per octave. The meaning of this, in physical terms, is that there exists a

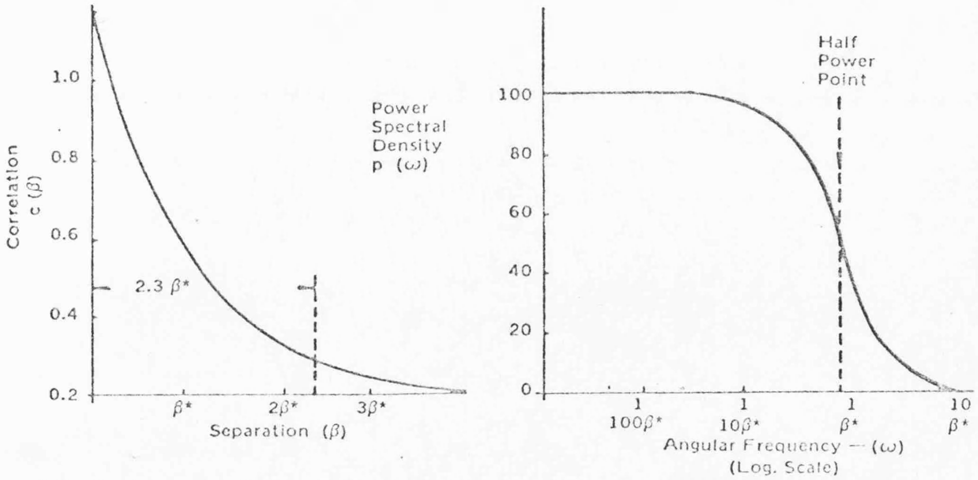


FIG. 3 THE MODEL: AUTOCORRELATION FUNCTION AND POWER SPECTRAL DENSITY.

wide band of lower frequencies (longer wavelengths) in the structure of the profile. At higher frequencies (shorter wavelengths) the magnitudes of the components present declines so that in this range the amplitude is proportional to the wavelength. It will also be noted that, at least in principle, these components continue to exist to an infinitely short wavelength.

Theory of Peak Distributions for the Random Signal Model

We now show how this random signal model of a surface profile can be transformed into a model which is appropriate to the theory of surface contact. It will be recalled that models of surfaces used in the theories of contact consist of a distribution of spherical asperities. This practice will be followed here, the characteristics of the asperities being derived from the characteristics of the peaks of the surface profiles. As in the earlier work on digital analysis of measured profiles, the peaks will be defined by three point analysis. An outline of this derivation will be given here; a more complete account will be published elsewhere [13].

The results obtained in this type of analysis depend upon the spacing of the three samples which are assumed to be separated by constant sampling intervals ℓ. Consider first samples separated by intervals $\ell = 2.3\beta^*$. It will be observed from equation (5)

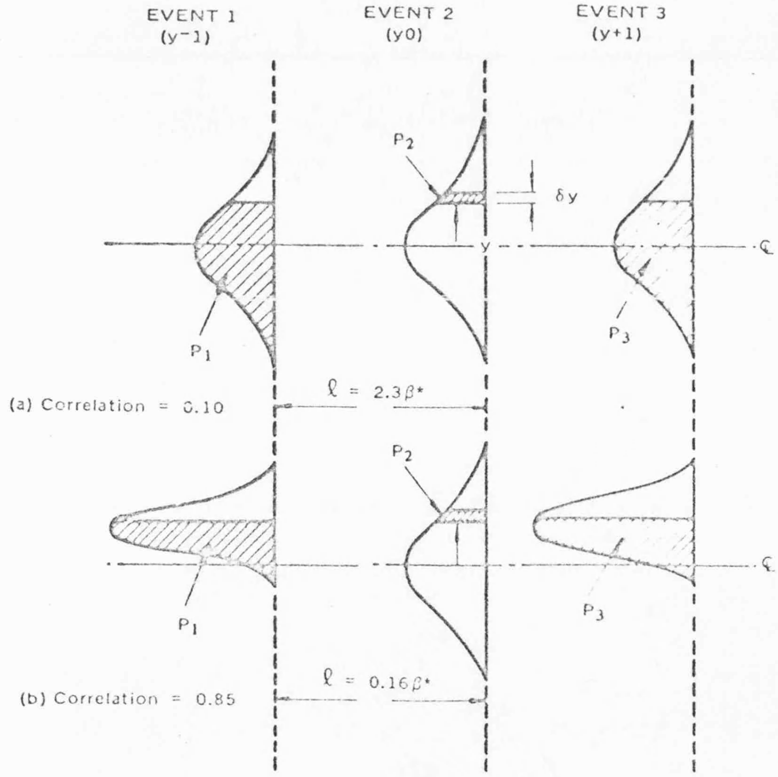


FIG. 4 MODEL USED IN DEDUCING DISTRIBUTION OF PEAKS.
(a) Sampling interval, $\ell = 2.3\beta^*$; correlation, $\rho = 0.10$
(b) Sampling interval, $\ell = 0.16\beta^*$; correlation, $\rho = 0.86$.

and Figure 3 that with this separation the correlation between successive events has declined to 0.10. This will be taken as the separation where the ordinates have just reached the condition where they can be regarded as independent events. The justification for this assumption is given elsewhere [13]. The probability distributions for the three ordinates are therefore each given by the original Gaussian distribution of height and this is shown in Figure 4(a). Thus the probability of the central ordinate being a peak at a height between y and $(y+dy)$ is given by three independent probabilities.

(a) the probability that event 1 (y_{-1}) is less than y
(b) the probability that event 2 (y_0) lies between y and $(y+dy)$
(c) the probability that event 3 (y_{+1}) is less than y.

The probability is therefore the product of these three probabilities given by the shaded areas $P_1 P_2$ and P_3 in Figure 4(a).

For shorter sampling intervals the two outer ordinates $(y_{+1}$ and $y_{-1})$ lie so close to the central point (y_0) that there exists a strong statistical correlation between the heights y_{-1} and y_{+1} and the central ordinate y_0. The effect of this is shown in Fig. 4(b) where it will be observed that the probability distributions of y_{-1} and y_{+1} are now modified by the fact that y_0 lies between y and $(y + \delta y)$. However, the probability that the central event is a peak between y and $(y + \delta y)$ is, once again, given by the product of the probabilities P_1, P_2, P_3 shown by the shaded areas of the distributions.

Figure 5 shows that similar techniques can be used to deduce the probability that a given ordinate is a peak of curvature C. Fig. 5(a) shows one combination of events which leads to this result when

$$C = 2y_0 - y_{+1} - y_{-1} \tag{7}$$

In this equation C is nondimensional but the true value of the curvature depends upon ℓ and is obtained by multiplying C by (σ/ℓ^2). Once again the probability is the product of the three shaded areas P_1, P_2, P_3 but the *total* probability that the central event is a peak of curvature C is obtained by considering all possible configurations which satisfy equation (7). Fig. 5(b) shows the similar system for shorter sampling intervals when a high correlation exists between successive events.

Using these methods the following expressions for the probability density functions have been derived. The probability density of an ordinate being a peak at height y is

$$f_p\,(y,\rho) = \frac{1}{4\sqrt{2\pi}}\left[1 + \operatorname{erf}\left(\frac{y}{\sqrt{2}}\sqrt{\frac{1-\rho}{1+\rho}}\right)\right]^2 \exp\left(-\frac{1}{2}y^2\right) \tag{8}$$

where ρ is the correlation between successive ordinates. For larger sampling intervals $\rho \to 0$, and equation (8) becomes

$$f_p\,(y) = \frac{1}{4\sqrt{2\pi}}\,[1 + \operatorname{erf}\,(y/\sqrt{2})]^2 \exp\left(-\frac{y^2}{2}\right) \tag{9}$$

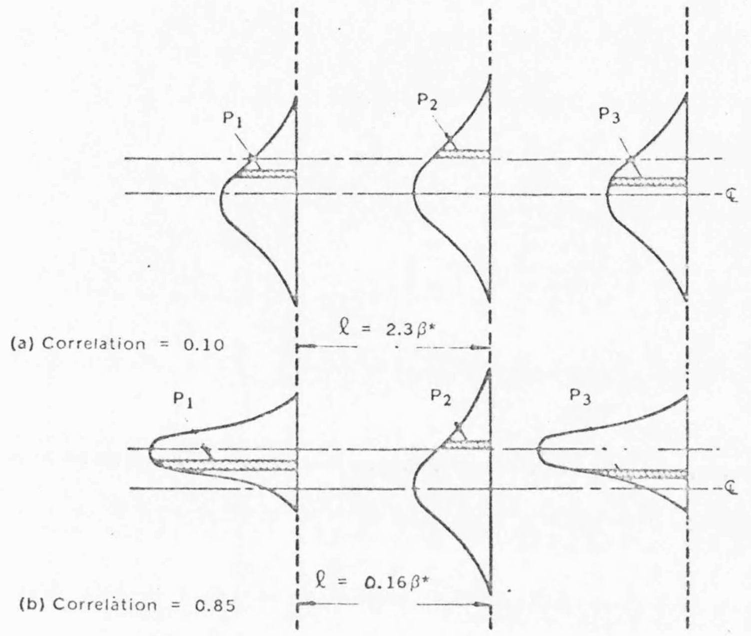


FIG. 5 MODEL USED IN DEDUCING THE DISTRIBUTION OF CURVATURES.
(a) Sampling interval, ℓ = 2.3β*; correlation, ρ = 0.10
(b) Sampling interval, ℓ = 0.16β*; correlation, ρ = 0.86.

Similarly the probability density of an ordinate being a peak of curvature C at height y is given by

$$f_c\,(y,C,\rho) = \frac{\exp\,(-\tfrac{1}{2}\,y^2)}{2\pi\sqrt{2\,(1-\rho^2)}}\,\exp\left[-\,\frac{[(1-\rho)\,y-\tfrac{1}{2}\,C]^2}{(1-\rho^2)}\right].\ \operatorname{erf}\left[\frac{C}{2\sqrt{1-\rho^2}}\right] \qquad (10)$$

As before, for large sampling intervals $\rho \to 0$ and equation (10) reduces to

$$f_c\ (y,C) = \frac{\exp\,(\tfrac{1}{2}\,y^2)}{2\pi\sqrt{2}}\,\exp\ [-\,(y-\tfrac{1}{2}\,C)^2]\ \ \operatorname{erf}\ [\tfrac{1}{2}C] \qquad (11)$$

The probability density of an ordinate being a peak of curvature C (at any height) is obtained by integrating equation (10) with respect to y. Thus

$$f_c\,(C,\rho) = \left[\frac{1}{4\pi\ (3-\rho)\ (1-\rho)}\right]^{\frac{1}{2}}\ \exp\left[\frac{-\,C^2}{4\ (3-\rho)\ (1-\rho)}\right]\operatorname{erf}\left[\frac{C}{2\sqrt{1-\rho}}\right] \qquad (12)$$

The ratio of peaks to ordinates N is

$$N = \frac{1}{\pi}\ \tan^{-1}\ \sqrt{(3-\rho)/(1+\rho)} \qquad (13)$$

The conclusions to be drawn from this analysis can be summarized as follows:

(1) In defining the functional characteristics of the profile by three point analysis or similar techniques the sampling interval used is a critical parameter.

(2) According to equations (8) and (9), if the height distribution is Gaussian the peak height distribution is also approximately Gaussian for all sampling intervals; Greenwood and Williamson [7] obtained this result from digital analysis of bead blasted surfaces using one sampling interval. Moreover, the characteristics of this peak height distribution depend upon the sampling interval and lie between two limiting conditions; these correspond either to a sampling interval so short that a high degree of correlation exists between successive samples, or, at the other extreme, to the sampling interval that will select a series of events which are effectively independent.

(3) Similarly the ratio of peaks to ordinates, N, depends upon the sampling interval. Equation (13) shows that at long sampling intervals ($\rho \rightarrow 0$) N is $^1/_3$ and at short sampling intervals ($\rho \rightarrow 1$) N is $^1/_4$. These limiting values can be explained in simple terms. $\rho \rightarrow 0$ implies that the three events are independent and the chance that any one of them (e.g., the central one) is the highest is obviously one third. As $\rho \rightarrow 1$ the modified distributions of y_{-1} and y_{+1} (Fig. 4(b)) become centered on the central y_0 so that the areas P_1 and P_3 become $^1/_2$ and the probability that the central event is a peak becomes $^1/_4$.

(4) The peak curvature distribution (equations (10) and (11)) is a function of the peak height. This dependence upon height increases as the sampling interval increases and the correlation between successive samples is consequently reduced.

(5) The mean curvature of the peaks increases with increasing peak height. For example, with a sampling interval of $2.3\beta^*$ (Fig. 4(a)), the curvature of the higher peaks is approximately three times that at the mean line.

(6) Theory and experiment show that the peak curvature distributions exhibit a general tendency to be skewed towards zero curvature. This has been previously demonstrated from digital analysis of profiles [7]. However our work shows that this distribution becomes more nearly Gaussian at the highest levels of the surface profile.

(7) The results given by Greenwood and Williamson [7] correspond, in the terms of our analysis, to an intermediate sampling rate. The characteristics of their peak height distribution (both its standard deviation and its disposition with respect to the height distribution) indicate a correlation co-efficient between successive samples of about 0.7.

(8) The methods described here can be extended in various ways. For example, definitions of a peak other than that described here can be accommodated. Similarly the same basic methods can be employed in the analysis of random profiles having height distributions other than Gaussian; where appropriate, graphical rather than analytic techniques can be used.

Digital Analysis of Profiles

We outline here the results of digital analysis of surface profiles and their comparison with the theory given above. The surface chosen, Aachen (64-13), is a typical ground surface used in an OECD cooperative research programme [14]. The results were obtained by converting the analogue signal from a Talysurf 4 instrument into digital form and subsequently processing the information in an ICL 1905 computer. A typical set of results used in the analysis consisted of some 10,000 ordinates with a sampling interval of $1\mu m$. By suitable selection of the results the effect of larger sampling intervals could be investigated. The Talysurf 4 instrument has a stylus which is a truncated pyramid, the dimension of the tip in the direction of traverse being approximately $2.5\mu m$; we term this the nominal tip dimension. A special series of experiments was also conducted using the Talystep apparatus using a special lightly loaded stylus having a nominal tip dimension of $0.25\mu m$. In this work the sampling interval was reduced to $0.25\mu m$.

Analysis of the results showed that the model used in this paper was a good representation of the data obtained from the surface. The distribution of heights was very close to Gaussian with an *rms* value (σ) of $0.5\mu m$ and the autocorrelation function was very close to exponential with a correlation distance (β^*, equation (5)) of $6.5\mu m$. Table 1 shows the values of the correlation (ρ) between successive samples for the values of the sampling interval (ℓ) used in this work.

Table 1
Relation Between Correlation (ρ) and Sampling Interval (ℓ) for Aachen 64-13

Sampling Interval (μm)	15	6.0	3.0	2.0	1.0	0.5	0.25
Correlation (ρ)	0.10	0.40	0.63	0.74	0.86	0.92	0.96

Figures 6 to 8 show comparisons between the theory, outlined above, and the results derived from the digital analysis of Aachen 64-13. Figure 6 shows the probability densities that an ordinate is a peak at a given height (Equation (8)) for sampling intervals of 15, 3 and $1\mu m$, corresponding to correlations of 0.10, 0.63 and 0.86 respectively. Similarly Figure 7 shows the probability density that an ordinate is a peak of curvature C (equation (12)) for these same sampling intervals. It will be observed that the agreement between theory and experiment is very good except at the shortest sampling interval.

This divergence between theory and experiment suggests that the measurements may be affected by the finite size of the stylus. Consider the results shown in Fig. 7. It will be observed that as the sampling interval is reduced one is concerned with the detection of peaks of larger and larger curvatures. In Fig. 7(c) a nominal value of the stylus curvature is indicated; this is taken as the reciprocal of the nominal tip dimension, as defined above. It will be seen that the divergence between theory and experiment occurs in the region expected if it were attributed to the finite size of the stylus.

The effect of stylus size is also shown in Fig. 8 in which the theoretical values of N, the ratio of peaks to ordinates (equation (13)), are compared with values derived from digital analysis of surface profiles. It will be recalled that according to the theory this ratio varies between values of $^1/_3$ for large sampling intervals $(\rho\rightarrow 0)$ and $^1/_4$ for very short sampling intervals $(\rho\rightarrow 1)$ Fig. 8 shows that for sampling intervals greater than $2\mu m$ $(\rho < 0.74)$ the ratio of peaks to ordinates detected by digital analysis of profiles agrees well with the theory. For shorter sampling intervals the numbers of peaks detected fall well below the theoretical values. Fig. 8 also shows that when using a stylus with smaller tip dimension upon the same surface the number of peaks detected in this region shows a significant increase.

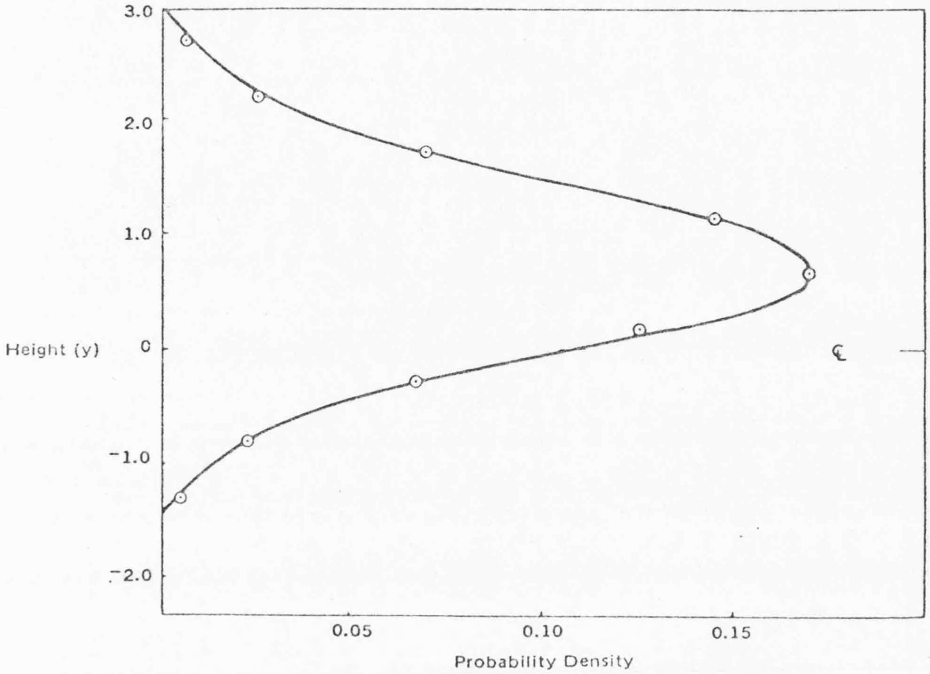


FIG. 6(a) PROBABILITY DENSITIES OF AN ORDINATE BEING A PEAK AT A HEIGHT *Y*
The full lines give the theory (equation 8). The experimental points are derived from digital analysis of profiles of Aachen 64-13. Results are shown for the values of the sampling interval (ℓ) corresponding to differing values of the correlation (ρ) between successive samples (a) ℓ = 15μm, ρ = 0.10

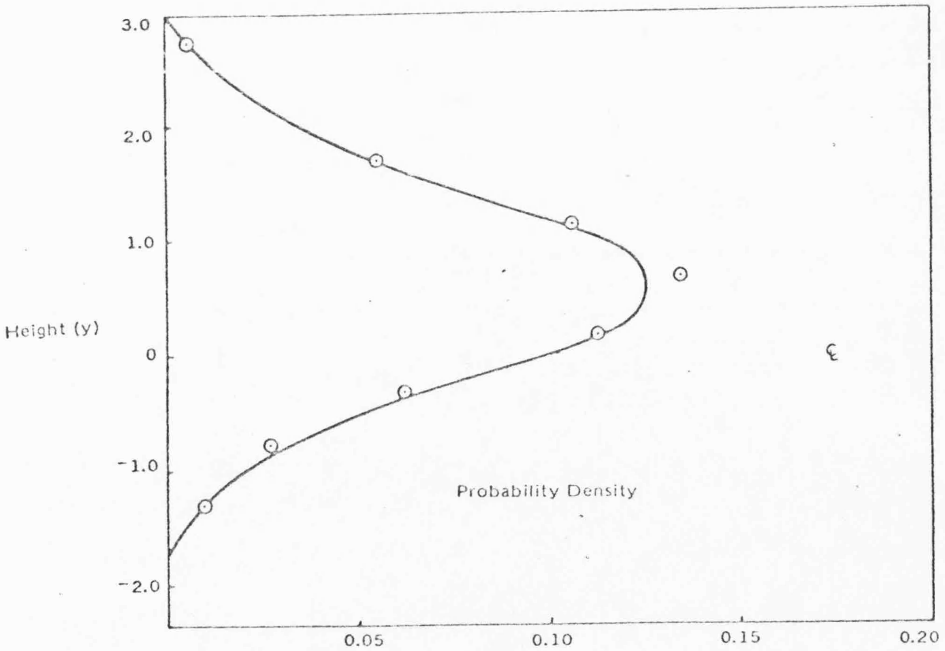


FIG. 6 (b) ℓ = 3.0μm, ρ = 0.63

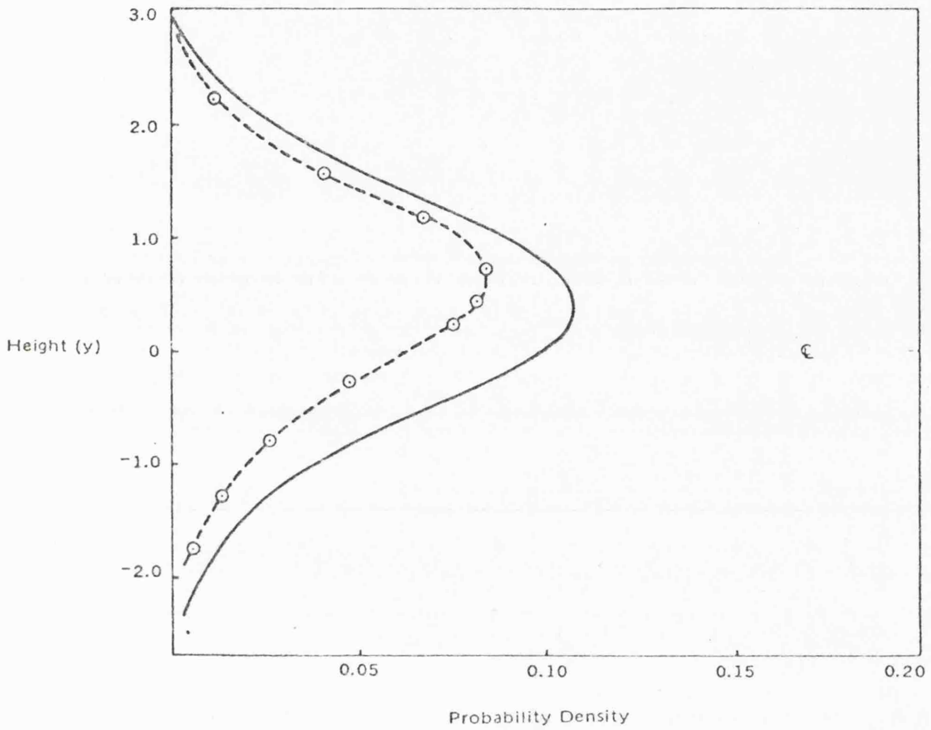


FIG. 6 (c) ℓ = 1.0μm, ρ = 0.86.

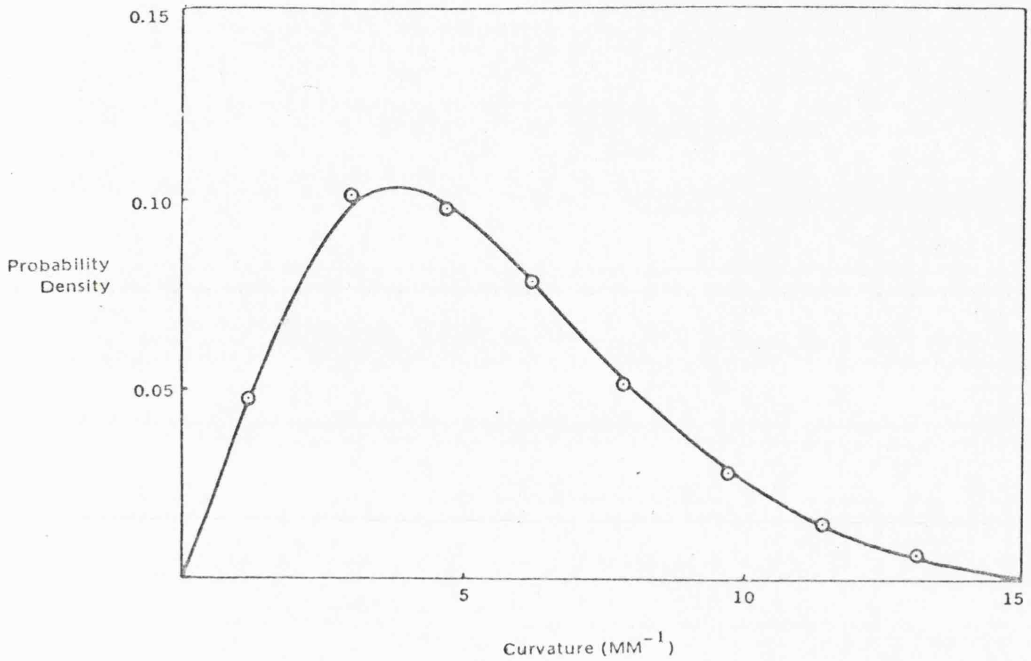


FIG. 7 PROBABILITY DENSITIES OF AN ORDINATE BEING A PEAK OF A GIVEN CURVATURE. The full lines give the theory (equation 12). The experimental points are from digital analysis of profiles from Aachen 64-13 (σ = 0.5μm, β^* = 6.5μm). Results are shown for three values of the sampling interval (ℓ) corresponding to differing values of the correlation (ρ) between successive samples. (a) ℓ = 15μm, ρ = 0.10

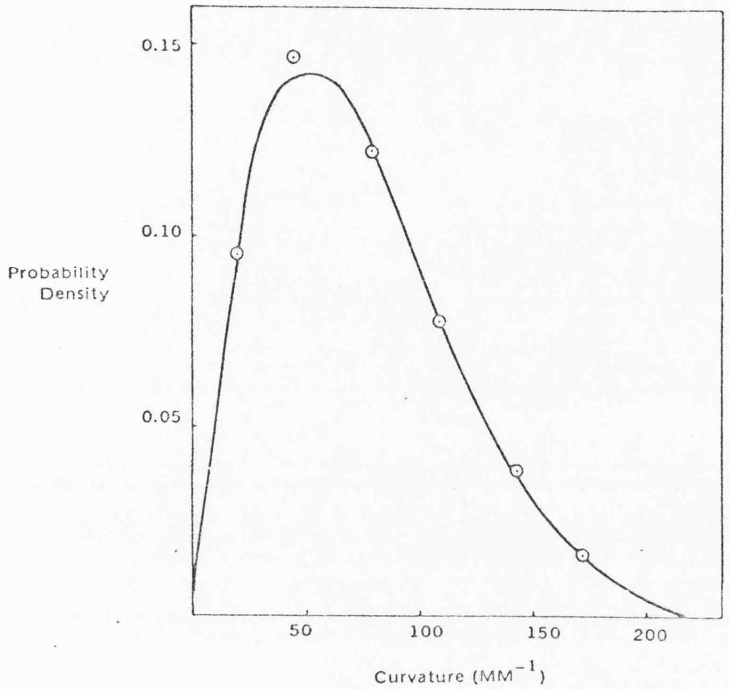


FIG. 7 (b) ℓ = 3.0μm, ρ = 0.63

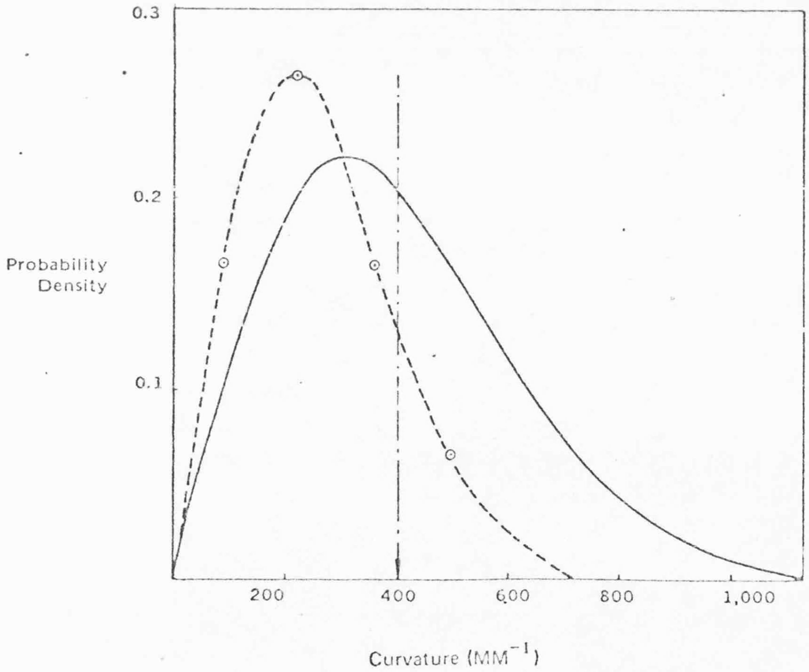


FIG. 7 (c) ℓ = 1.0μm, ρ = 0.86.

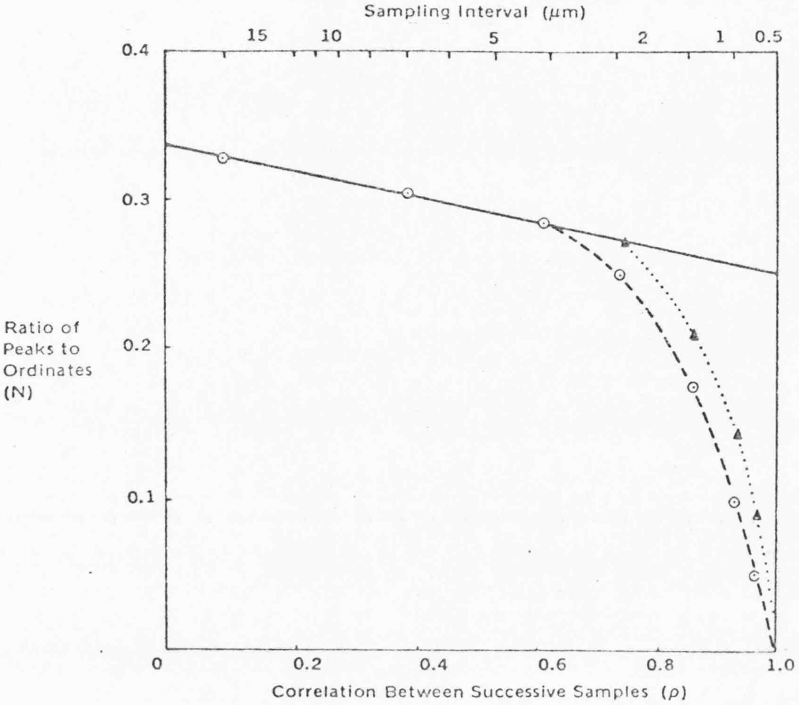


FIG. 8 RATIO (N) OF PEAKS TO ORDINATES AS A FUNCTION OF THE CORRELATION (ρ) BETWEEN SUCCESSIVE SAMPLES. The full line gives the theory (equation 13). The experimental points are derived from digital analysis of profiles from Aachen 64-13.
Using normal stylus, nominal tip dimension 2.5μm.
Using special stylus, nominal tip dimension 0.25μm.

This comparison between the theory and the results of digital analysis of surface profiles shows that stylus instruments are capable of detecting surface features covering two orders of magnitude (see the magnitudes of curvatures in Fig. 7) and this covers most of the features of surface structure known to be of functional significance; at the smallest scale of size, however, the resolution may be influenced by the finite size of the stylus. There is some evidence that features of the stylus, other than the tip dimension, influence the resolution [13] and, clearly, this is a subject worthy of further study.

An Experimental Study of Surface Profiles During Running-In

We now discuss some experiments which form part of a larger programme devised to study the factors of surface topography of significance in friction and wear. The experiments to be described were conducted using a simple crossed cylinders friction machine [15] in which two cylindrical specimens were loaded together and rubbed at low sliding speeds. The direction of sliding (Fig. 9) was at 45° to the axes of the two cylinders so that, during sliding, the point of contact moved constantly to a new region of both the specimens.

In these experiments arrangements were made to locate one of the specimens accurately, both in the friction machine and in the Vee-block of the Talysurf apparatus. In this way it was possible to measure the profile of *the same track* on the specimen, both in its original state and after one or more traversals of the loaded contact in the friction machine. As a demonstration of the accuracy of the relocation techniques Fig. 10* shows profiles (both examples repeated after relocation) of two parallel tracks on the surface of a specimen separated by a distance of 5×10^{-4} in $(1.27 \times 10^{-3}$ cm).

In the experiments to be described the cylindrical specimens were 0.25 in (0.635 cm) diameter and were of a 0.5% plain carbon steel hardened and tempered to a hardness of 300 DPN. They were rough ground to a finish of 32×10^{-6} in *c/a* and were rubbed together under conditions of boundary lubrication using a plain mineral oil as a lubricant.

Figure 11(a) shows the profile of a specimen before running and Figure 11(b) shows the profile of the same track on the specimen after one traversal under a load of 2.5 Kg. It will be observed that as a result of one traversal of the load the tops of the higher asperities have been greatly modified by plastic deformation but the valleys are largely unaffected. This is shown in Fig. 11(a) where the modified profile has been superposed on the original profile. Where the two profiles differ the modified profile has been shown by a broken line. This same technique has been adopted in Fig. 12 which shows the original profile of a similar specimen and the superposed profiles obtained after a series of experiments involving successive traversals at loads of 0.25, 2.0 and 8.0 Kg.

*These profiles involve no rubbing in the friction machine and were devised simply to demonstrate the accuracy of the re-location system. We are indebted to Mr. M. S. Lunn for these results which were undertaken as part of a final year undergraduate project in the Department of Engineering of the University of Leicester.

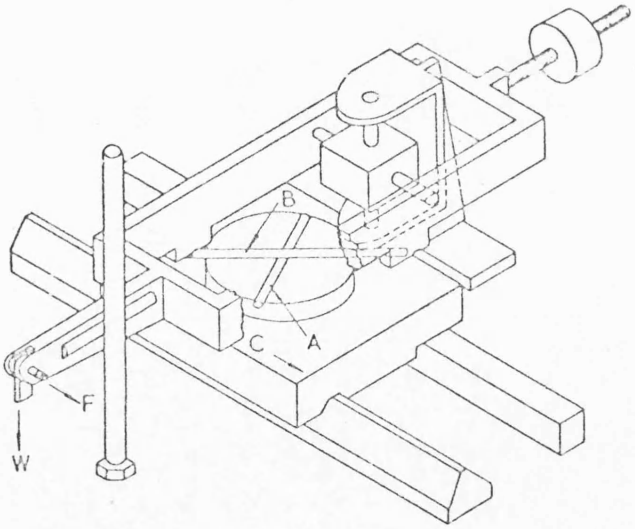


FIG. 9 THE CROSSED-CYLINDERS FRICTION MACHINE.
A, Lower specimen. B, upper specimen. C, direction of motion of lower specimen and carriage. W, load. F, calibration of frictional force.

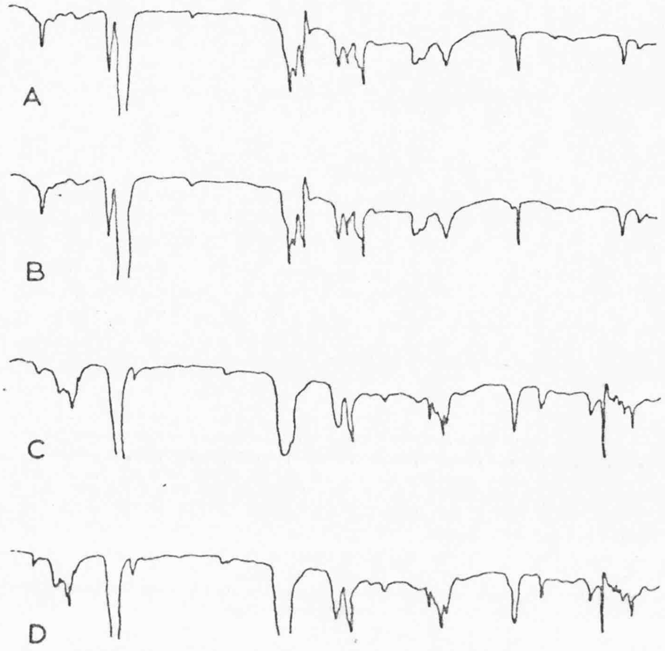


FIG. 10 DEMONSTRATION OF THE ACCURACY OF RELOCATION.
Records A and B are taken along the same track, relocating the specimen and re-adjusting the micrometer between the two recordings. Records C and D are taken along a parallel track 0.0005 in (1.27×10^{-3} cm) from that of A and B; as before, specimen relocated and micrometer adjusted between tracks.

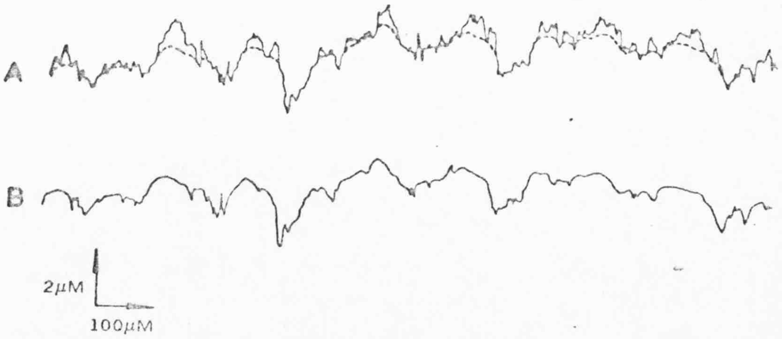


FIG. 11 TALYSURF PROFILES OF CYLINDRICAL SPECIMENS USED IN LUBRICATED FRICTION EXPERIMENTS. (A) Original surface profile (with profile of (b) superposed). (B) Profile of the same line after one traversal of the load. Specimens 0.5% C. steel, 300 D.P.N., 0.635 mm diameter, cla 32×10^{-6} in. Load 2.5 Kg.

Figures 11 and 12 show that one immediate effect of rubbing is that much of the shorter wavelength structure of the surface finish at the regions of contact is lost as a result of the first traversal of the load. The dominance of the longer wavelength structure of the surfaces after rubbing is also illustrated in Fig. 12 by reference to the rolling circle envelope. Above each profile in Fig. 12 is displayed this rolling circle envelope corresponding to the locus of the lowest point of a 0.25 in (0.635 cm) diameter sphere (representing here the upper specimen) moved over the original profile without deformation. This technique operates as a mechanical filter [8] and the resemblance between the rolling circle envelope and the profiles of the surface after rubbing will be noted.

The fact that these changes in the profile are produced in the first few traversals is demonstrated in Fig. 13. As a simple measure of changes in the surface topography, the change in the *cla* value of the roughness is used. This change is plotted as a function of

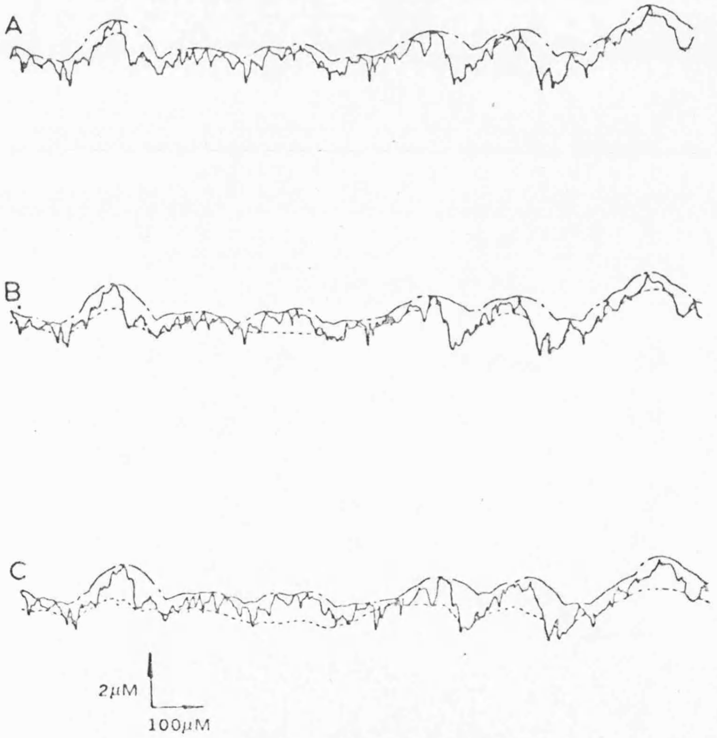


FIG. 12 TALYSURF PROFILES OF SPECIMENS USED IN LUBRICATED FRICTION EXPERIMENTS. In this experiment three traversals of the load were made and the loads used in successive runs were 0.25 Kg, 2.0 Kg and 8.0 Kg. Each record shows the original profile and the profile at the particular stage of the experiments. Above the profiles is also shown the rolling circle envelope (see text).
(A) Profile after one run (B) Profile after two runs (C) Profile after three runs
The conditions were otherwise as in Figure 11.

the number of repeated traversals at the same load. It will be seen that after the first few traversals further changes in roughness are very small indeed.

Discussion

It has long been recognized, in the study of random processes, that a random signal can be completely defined, in a statistical sense, by two parameters: the height distribution and the autocorrelation function. The present paper takes this same model for the profile of a random surface. In the particular model used in this paper these two parameters become simply two lengths associated with the y and x coordinates of the profile; these lengths are the standard deviation of the height distribution (σ) and the correlation distance (β^*). It can be shown [13] that all the required geometric parameters of the profile can be defined in terms of these two independent parameters. Thus it is possible to derive statistical distributions of peak heights, peak curvatures and slopes. Although many surfaces do not conform exactly to the model of this paper, nevertheless the principle of characterizing surface finish in terms of two such independent parameters seems capable of wider application. In practice, for reasons connected with ease of measurement, it may be desirable to measure a derived parameter such as the mean slope.

The theory and its comparison with the results of digital analysis of surface profiles also raises a fundamental problem associated with the representation of surfaces by models. The random signal model is a complete description of the profile in a statistical rather than a deterministic sense. In transforming this model into a form suitable for the study of surface contact the techniques of three point analysis were used. It was then found that the results thus derived depend upon the sampling interval used. The meaning of this result is that three point analysis causes, inevitably, a loss of information compared with that which was present in the random signal model or that which existed in the original surface profile. Three point sampling provides information about that part of the surface

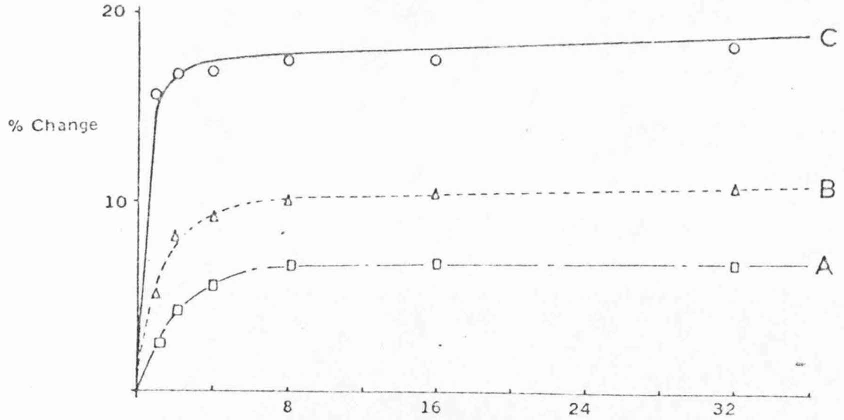


FIG. 13 CHANGE IN SURFACE ROUGHNESS AFTER REPEATED RUNS.
Repeated runs were made at a constant load the surface finish being measured throughout the series of experiments which were performed for three different loads. (A) 0.2 Kg, (B) 1 Kg, (C) 2.5 Kg.
The graph shows the percentage change in *cla* roughness with the number of runs. Experimental conditions otherwise as in Fig. 11.

structure which has a wave length comparable with the sampling interval. An over-emphasis upon the limitations of three point analysis is perhaps irrelevant here; what is more important is a broader physical interpretation of the results of the analysis.

Consider therefore the distribution of peak heights (equations (8) and (9) and Fig. 6). The results obtained can first be considered in terms of two extremes. It has been a theme of this paper that a profile can be considered as a series of events, independent when widely spaced, highly correlated when closely spaced. When the profile is considered as a series of closely spaced events this means that one is concerned with the finest scale of structure. Thus, in the limit, the finest scale of structure has a peak distribution which is identical with the overall distribution of ordinates, adding the proviso, argued above, that only one in every four events is a peak. This emphasizes the point that the profile consists of superposed asperities of a wide range of sizes. The model adopted here is, therefore, in a very reai sense, analogous to the model of Fig. 1(c). Figure 1(c) is a deterministic model with superposed asperities of three different scales of size. We have shown that the random signal model of this paper consists of asperities covering (at least in principle) an infinitely wide range of sizes superposed in a random fashion and which can be interpreted only in a statistical sense. It is from the complexity of this type of structure that the difficulties associated with the theory of surface contact arise.

The results of the experiments, shown in Figs. 11 and 12, suggest that there is a main, broad scale, structure which is of the greatest significance in surface contact. It has been argued elsewhere [13] that this structure corresponds to a wavelength of approximately $10\beta^*$. The theoretical arguments for a scale of this order of size may be summarized as follows:

(a) Events separated by a distance $2.3\beta^*$ have a correlation of 0.1 and can be regarded aş those which have just reached the condition where they can be regarded as effectively independent.

(b) The components of the spectrum of greatest power (Fig. 3(b.) are in a band of wavelengths with an upper limit of about $2\pi\beta^*$.

(c) The mean distance between mean line crossings [16] (after removing wavelengths shorter than $2\pi\beta^*$) is $5.5\beta^*$.

This emphasis upon answering the question of what constitutes the main structure of practical random profiles which contain a wide range of wavelengths is of importance in considering the practical aspects of surfaces in contact. It is now recognized that in many, in perhaps the vast majority, of the asperity contacts the deformation is entirely elastic. It therefore seems possible that many forms of surface failure may be linked with the incidence of plastic deformation. Two criteria had been suggested which link the probability of plastic deformation to the topography of the surface and the mechanical properties of the material. Blok [17] and Halliday [18] have suggested a criterion which is based upon a calculation of the shape of an asperity which can be pressed flat into the surface without plastic flow. This can be cast in the form

$$\bar{m} \leqslant K\left(\frac{H}{E'}\right) \tag{14}$$

where $\bar{m}$ is the average slope and K is numerical constant, in the range 0.8 to 1.7, which depends upon the assumed shape of the asperity. Greenwood and Williamson [7] used a model of the surface similar to Fig. 1(a) in which the asperities, all of the same radius of curvature R, were disposed in a Gaussian distribution of heights having a standard deviation σ_p. It was shown that the probability of plastic flow depended very little upon the load but was critically dependent upon a plasticity index ψ given by

$$\psi = \left(\frac{E'}{H}\right)\left(\frac{\sigma_p}{R}\right)^{1/2} \tag{15}$$

Using the random signal model of the present paper both of these criteria take the same form. The Blok-Halliday criterion can be expressed as

$$\bar{m} = \gamma_1\left(\frac{\sigma}{\beta^*}\right) \leqslant K\left(\frac{H}{E'}\right) \tag{16}$$

and the plasticity index of equation (15) takes the form,

$$\psi = \gamma_2\left(\frac{E'}{H}\right)\left(\frac{\sigma}{\beta^*}\right) \tag{17}$$

where, in these equations, γ_1 and γ_2 are numerical constants depending upon the scale of size used. The derivation of equations (16) and (17) can be readily understood since the theory shows that the slopes on a random surface are proportional to σ/β^*, σ_p is proportional to σ and R is proportional to $(\beta^*)^2/\sigma$. However, a more careful consideration of the way in which these plasticity criteria are derived shows the difficulties involved. They are based upon the assumption that the deformation of each asperity can be regarded as independent. Therefore for the deterministic models of Fig. 1 they are valid only for the simpler model of Fig. 1(a). Similarly for the random signal model of this paper plasticity criteria are valid only if they are applied to the main, long wavelength, structure; they then show the probability of plastic flow over regions associated with this scale of size. Using the equations corresponding to a sampling interval, $\ell = 2.3\beta^*$ it can be shown [13] that in equations (16) and (17) $\gamma_1 = 0.24$ and $\gamma_2 = 0.3$.

Finally, we return to a major theme of this paper; the concept that random surfaces are defined by two parameters; σ and β^*. Consider the way in which such random surfaces are generated by mechanical methods. These processes involve multiple contacts between particles and the surfaces; for example, in grinding and sand blasting the unit event is the interaction between a grit and the surface resulting in the removal or displacement of material. If one postulates a random element in this interaction, it seems likely that the surfaces thus generated could well have a random structure in which the value of σ bears a simple relationship to the depth of the unit event and β^* bears a simple relationship to the width of the unit event. Thus a characterization of typical random surfaces in terms of σ and β^* is not only a valid description of their functional behavior but also may be closely linked with the mechanism of their generation.

Acknowledgment

We are indebted to the Directors of Rank Precision Industries Ltd., for financial support and for permission to publish.

References

[1] Bowden, F. P., and Tabor, D. "The Friction and Lubrication of Solids." Oxford, *University Press,* 1950.

[2] Timoshenko, S., and Goodier, J. N. "Theory of Elasticity." *McGraw-Hill Book Co.,* New York, 1951.

[3] Tabor, D. "The Hardness of Metals." Oxford, *University Press,* 1951.

[4] Archard, J. F. "Elastic Deformation and the Laws of Friction." *Proc. Roy. Soc. Lond.,* Vol. 243 Series A, 1957, p. 190.

[5] Archard, J. F. "Single Contacts and Multiple Encounters." *J. Appl. Phys.,* Vol. 32, 1961, p. 1420.

[6] Greenwood, J. A., and Williamson, J. B. P. "Contact of Nominally Flat Surfaces." *Proc. Roy. Soc. Lond.,* Vol. 195 Series A, 1966, p. 300.

[7] Greenwood, J. A. "The Area of Contact Between Rough Surfaces and Flats." *Trans. ASME,* Vol. 89 Series F., (J. Lubric. Techn.) 1967, p. 81.

[8] Reason, R. E. ' Workshop Requirements of Surface Measurement." *Proc. Instn. Mech. Engrs.,* Vol. 182 Part 3K (Conference on the Properties and Metrology of Surfaces) 1967-68, Paper No. 23, p. 300.

[9] Peklenik, J. "New Developments in Surface Characterization and Measurements by Means of Random Process Analysis." *Proc. Instn. Mech. Engrs.,* Vol. 182 Part 3K (Conference on the Properties and Metrology of Surfaces) 1967-68, Paper No. 24, p. 108.

[10] Whitehouse, D. J., (To be published)

[11] Bendat, J. S. "Principles and Applications of Random Noise Theory." *John Wiley,* New York, 1958.

[12] Beckmann, P., and Spizzichino, A. "The Scattering of Electromagnetic Radiation from Rough Surfaces." *Pergamon Press,* London, 1963.

[13] Whitehouse, D. J., and Archard, J. F. (To be published)

[14] OECD. Sub-group Typology and Topology (To be published)

[15] Archard, J. F. "A Crossed-Cylinders Friction Machine." *Wear,* Vol. 2, 1958-59. p.21.

[16] Rice, S. O., see Bendat, J. S. loc. cit. p. 128.

[17] Blok, H. "Comments on a Paper by R. W. Wilson." *Proc. Roy. Soc. Lond.,* Vol. 212 Series A, 1952, p. 480.

[18] Halliday, J. S. "Surface Examination by Electron Microscopy." *Proc. Instn. Mech. Engrs.,* Vol. 169, 1955, p. 777.

Paper 12

IMPROVED TYPE OF WAVEFILTER FOR USE IN SURFACE-FINISH MEASUREMENT

By D. J. Whitehouse*

This paper is concerned with the use of phase-corrected filters in surface-texture measurement. It is divided into two parts; Part 1 is devoted to the requirements of surface-texture measurement, and Part 2 is concerned with the relevant theory.

Part 1—Application

INTRODUCTION

A NECESSARY PRELIMINARY to numerical assessment of surface profiles is to extract the frequency components representative of the undulations to be measured (the so-called roughness) and to eliminate those that would be irrelevant—in particular, the general slope and curvature of the surface relative to the mechanical datum of the instrument.

This separation can be achieved graphically by restricting the length of sample used in the assessment of the roughness to a given length called the sample length. Assessment of a number of such sample lengths placed end to end usually gives a fair estimate of the surface roughness. A simple and convenient method to achieve this separation instrumentally is to pass the profile signal coming from the pick-up through an electric wavefilter before applying it to a meter. Many early instruments made use of a filter which had a transmission characteristic resembling that of two capacitor-resistor networks connected in cascade in such a manner that the second did not load the first, the cut-off wavelength of the filter being made nominally equal to the sample length. Eventually this type of filter, with specified cut-off wavelengths, was standardized in the American–British–Canadian Standards. It will be referred to in this paper as the standard filter.

For industrial purposes the standard filter has considerable merits. It is simple and has the property of establishing a mean line naturally which does not suffer from the discontinuities sometimes present in the construction of the mean line according to the accepted graphical procedure (1)† (2). Moreover, this mean line is established without dependence on instrument adjustment. It may also be mentioned that the mean line of the filter can now be calculated directly using equations and a technique developed at the Rank Taylor Hobson Research Laboratory (3). Because of its simplicity, its very widespread use, and the fact that it is demonstrably serviceable for a great many applications, the standard filter seems likely to find continued industrial acceptance for many years to come.

Nevertheless, the standard filter has the disadvantages of a rather gradual transition from the nominal 'pass' band of the filter to the nominal rejected region (Fig. 12.1*a*), and of distorting the signal within the pass band (sometimes appreciably) because of phase shift. These disadvantages set problems in standardization and calibration of instruments, and also make it difficult to assess some of the parameters of the roughness when undulations with wavelengths near to the cut-off are present.

It is the purpose of this paper to discuss the possibility of a filter which introduces no phase distortion and has either the same or a more suitable amplitude transmission characteristic than the standard filter. Such a filter is shown to be feasible and highly desirable, but it would be more costly at present than the standard filter.

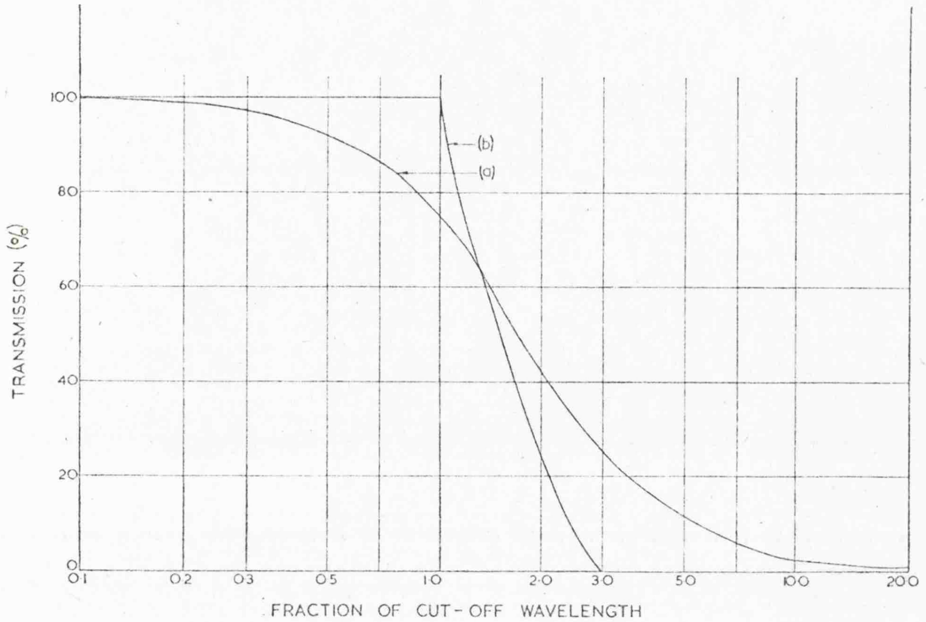
The standard wavefilter—phase-corrected wavefilter

The amplitude transmission characteristics shown in

The MS. of this paper was received at the Institution on 15th January 1968.

* *Rank Taylor Hobson, Leicester.*

† *References are given in Appendix 12.1.*



a Standard wavefilter characteristics. *b* 3:1 phase-corrected or linear phase filter.

Fig. 12.1. Comparison between standard and phase-corrected filters

Fig. 12.1*a* for the standard wavefilter are defined in British Standard 1134 and American Standard B46. Seventy-five per cent of the amplitude of those undulations whose wavelength is equal to the accepted cut-off is transmitted by the filter. Also a maximum rate of attenuation of 12 dB/octave is specified (Fig. 12.1*a*). The signal emerging from the filter is often referred to as a filtered or modified profile. It is usually assessed in surface-measuring instruments by measuring one or other of its parameters, such as the c.l.a. or r.m.s. value.

Before proceeding to discuss these characteristics, it is useful to imagine that any profile comprises a number of sinusoidal components each of different amplitude and wavelength. It is not quite so simple for purely random waveforms, but the picture is still useful. When the profile is passed through the standard wavefilter, each of these components gets attenuated according to its wavelength. Also, these components get shifted in time by different amounts depending upon the phase–wavelength characteristic of the filter, so that the filtered profile does not represent only the simple subtraction of the attenuated components from the original profile, but is also modified by the relative shifting of the transmitted components. The result of this relative shifting is that the filtered profile is a distorted version of the original profile. This distortion is called phase distortion.

A point to note is that an electric wavefilter reacts only to the way the input signal behaves in time. Consequently, to find out how the filter deals with a profile it is necessary to know how the profile is converted to a time-dependent signal. Actually this is very simple; the measuring instrument incorporates a scanning device such as a motor and gearbox which causes a stylus to track across the surface at a given speed, say v in/s. In this way an undulation of λ in wavelength on the surface is transformed into a frequency of v/λ in time.

In order to elucidate the nature of possible improvements, the manner in which the standard filter responds to some idealized repetitive waveforms will be considered first. This is because the standard wavefilter is generally satisfactory for random roughness waveforms. Some practical cases will be considered later. Repetitive waveforms tend to accentuate any shortcomings.

Look first at Fig. 12.2 which shows how the standard filter responds to sine waves. The figures show sine waves of different wavelengths. The full line shows the original wave, the dotted line is the mean line found by the wavefilter, this being the line that accounts for the output behaviour. The difference at any instant between the two lines is the filtered profile value. The top four graphs are within the pass band; notice that the mean line is displaced to the right relative to the original sine wave in all cases, which means that it is delayed in time relative to the profile, and further that the mean line has a large amplitude of undulation even for wavelengths much smaller than the cut-off, e.g. for one-third of the cut-off (the top graph). Another point is that the amplitudes of the mean line can be of the same order as the original sine wave, even just outside the pass band. Despite these points the filtered profile, which is the difference at any position between the mean line and the profile, obeys exactly the specified amplitude transmission characteristic mentioned previously. As an example, at the cut-off the mean line has an amplitude of just over 90 per cent of the original sine wave profile, but the filtered profile has the correct amplitude of 75 per cent of the amplitude of

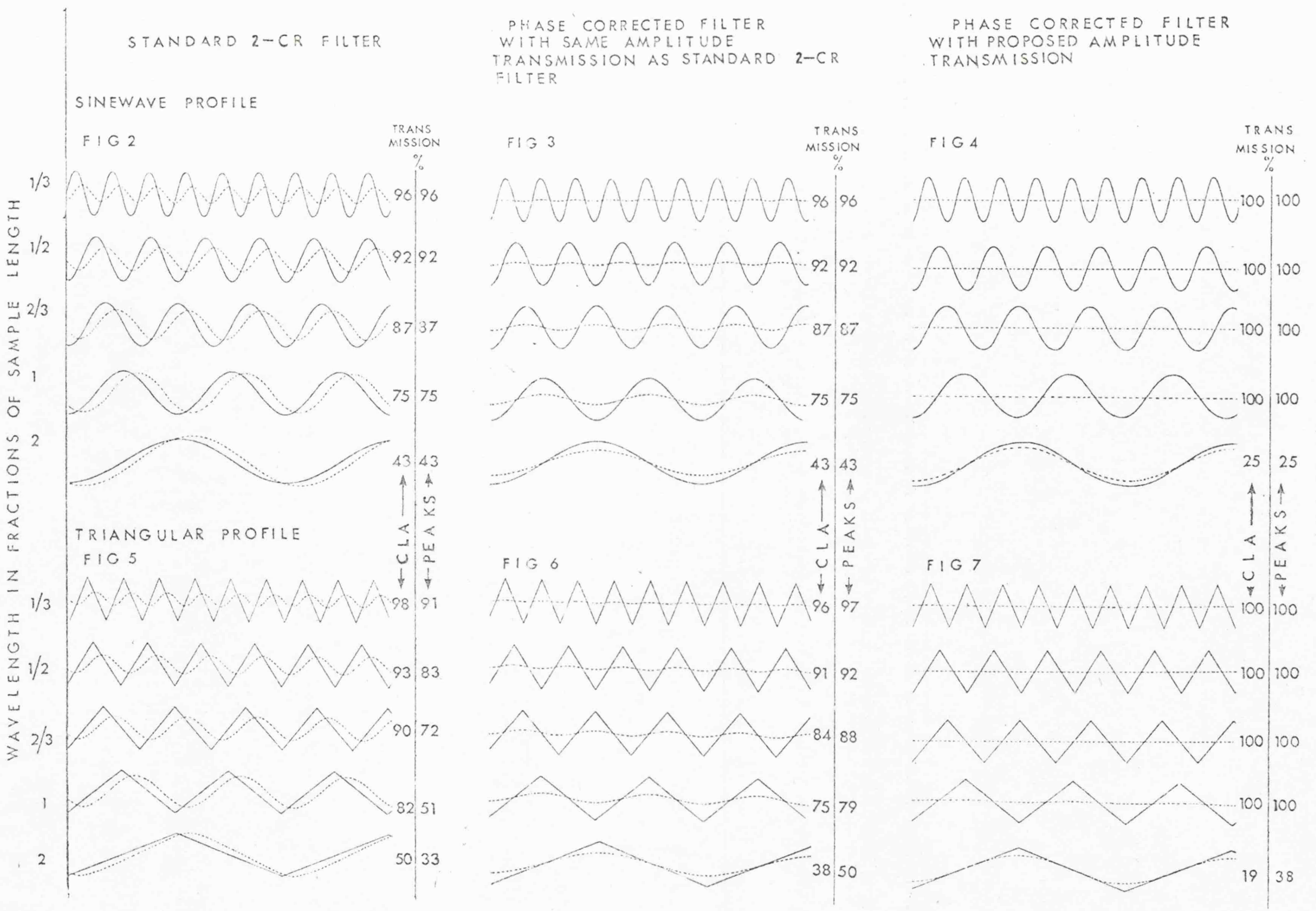


Fig. 2. Sine wave input to standard wavefilter

Fig. 3. Sine wave input to phase-corrected standard wavefilter

Fig. 4. Sine wave input to 3:1 phase-corrected filter

Fig. 5. Triangular wave input to standard wavefilter

Fig. 6. Triangular wave input to phase-corrected standard filter

Fig. 7. Triangular wave input to 3:1 phase-corrected filter

Figs 12.2–12.7. Comparison of standard and phase-corrected filters the profile. The amount by which the mean line is shifted relative to the profile depends on the wavelength of the sine wave profile.

For most commonly used filters, specifying the amplitude transmission characteristics automatically fixes the phase characteristics. This class is known as minimum phase, of which the standard filter of two CR networks is a typical member. Specifying the transmission characteristics for the standard wavefilter automatically implies a phase-shifted mean line.

Suppose that a filter is available which has the same transmission characteristic as the standard filter, but at the same time has a mean line which is not shifted in phase relative to the profile. How is the mean line amplitude affected?

This situation is shown in Fig. 12.3 for the same sine waves as in Fig. 12.2, and it can be seen that the mean line is nearly straight at one-third of the cut-off. Compare this with the same profile for the standard wavefilter. There is a dramatic reduction in the amplitude of the mean line, and this is true for all the cases shown. At the cut-off the filtered profile for this new filter has the required amplitude of 75 per cent of the profile—the 25 per cent attenuation is accounted for entirely by the amplitude of the mean line. In other words, if the mean line is kept in phase with the sine wave profile, then for any value of wavelength the maximum amplitudes of the filtered profile and the mean line add up to unity—a criterion which does not hold in the case of the phase-shifted mean line. A direct result of this is that the mean line in phase with the profile undulates less. Summarizing, it may be said that the mean line becomes straight much nearer to the cut-off in the filter whose mean line is in phase than it does in the standard wavefilter, although the filters have precisely the same amplitude transmission characteristics. This fact is of fundamental importance.

For a sine wave profile the phase distortion simply takes the form of a phase shift of the filtered profile relative to the original profile, but for any other profile which can be considered to be made up of a number of such sine wave components, the distortion is more complicated.

Consider now Figs 12.5 and 12.6. These show triangular waveform profiles of differing wavelengths. Remembering that the filtered profile is the difference at any point between the mean line and the profile waveform, it can be seen that the filtered profile for the zero phase-shifted mean line bears a much closer resemblance to the original waveform than it does for the standard wavefilter.

The zero phase filter has a more realistic mean line because the sine wave components making up the triangular waveform are not shifted relative to each other in the filter. Consequently, the components have the same relative positions upon emerging from it. Hence, even taking account of those components that have been attenuated in the filter, the output still resembles the input. This is not so true in the case of the standard wavefilter. Distortion of the filtered profile can make it difficult to assess numerically. This is the problem that can be encountered in practice when highly repetitive profiles just within the pass band are put into the standard wavefilter. As an example, the triangular waveforms shown in Figs 12.5 and 12.6 are a close enough approximation to a practical waveform to illustrate how the problem arises. Consider the filtered profile in Fig. 12.5; the concept of a peak output at the cut-off, say, is difficult to imagine—the peak shape has virtually disappeared. Now look at Fig. 12.6 where the peak is noticeable and unambiguous to measure.

So far only the phase characteristics have been considered. It is also desirable that the mean line should be straight for all wavelengths within the pass band of the filter, i.e. up to the cut-off. This would mean that the profile would suffer no attenuation for wavelengths up to the cut-off. From a surface-roughness measurement point of view this seems sensible, because it is natural to suppose that if a cut-off has been chosen for the filter of a value larger than the longest roughness wavelength on the surface, then all the roughness will be passed unattenuated. Another point about the transmission characteristics which can be mentioned concerns the behaviour outside the cut-off. Although the behaviour outside the cut-off is not so important as that within the cut-off it still has some relevance to the measurement of the roughness. The standard wavefilter tends to fall off too gradually for wavelengths longer than the cut-off, with the result that waviness components can be included in the roughness assessment.

From these factors it appears that an amplitude transmission characteristic which is unity up to the cut-off wavelength and which falls rapidly after the cut-off would be more suitable for roughness measurement. Excessively high rates of attenuation, however, could be unrealistic mechanically because a considerable variation is not expected in the functional behaviour of, say, two surfaces having roughness of equal amplitude but slightly different wavelength. Fig. 12.1*b* shows one such characteristic that has seemed practical, having unity transmission up to the cut-off and then falling off to zero at three times the cut-off, the rate of attenuation being linear with equivalent frequency.

Figs 12.4 and 12.7 show how a filter having an in-phase mean line similar to the one mentioned previously, but having the new transmission characteristics, behaves to the sine and triangular waveforms. The figures show a straight mean line right up to the cut-off and no distortion of the filtered profile. This is what could justifiably be called a filter with a well-behaved mean line. This filter will be referred to as the phase-corrected filter.

So far only idealized waveforms have been shown. In fact these accentuate the difficulties encountered in practice. Mean lines, as a rule, do not oscillate with such a large amplitude for practical waveforms within the cut-off even for periodic profiles, because a random component is always present. In the case of profiles which are random, the undulation of the mean line and distortion of the filtered profile from the standard wavefilter are not so obvious. The majority of profiles of this type have realistic mean lines, providing that the longest spacings are short compared with the cut-off length. However, for the standard wavefilter the distortion of the filtered repetitive profile and the undulation of the mean line do present a serious enough problem in some instances to warrant correction.

Some practical profiles are shown in Figs 12.8 and 12.9 together with a comparison of the mean line found by the standard wavefilter and the phase-corrected wavefilter. They show that the phase-corrected filter has advantages over the standard wavefilter in a number of ways. The gain is usually greatest when other factors apart from roughness measurement have to be considered, such as only a small amount of available traverse or waviness components lying near to the dominant tool marks on the surface. In fact, the advantage is usually greatest in cases where it is impossible for one reason or another to choose a cut-off for the standard filter long enough to make phase distortion negligible. Many of the parameters used to measure roughness such as the peak value, the derivatives, or the bearing ratio curve are affected by phase distortion. Centre-line average is affected to a lesser extent, while r.m.s. and some related parameters are not affected at all.

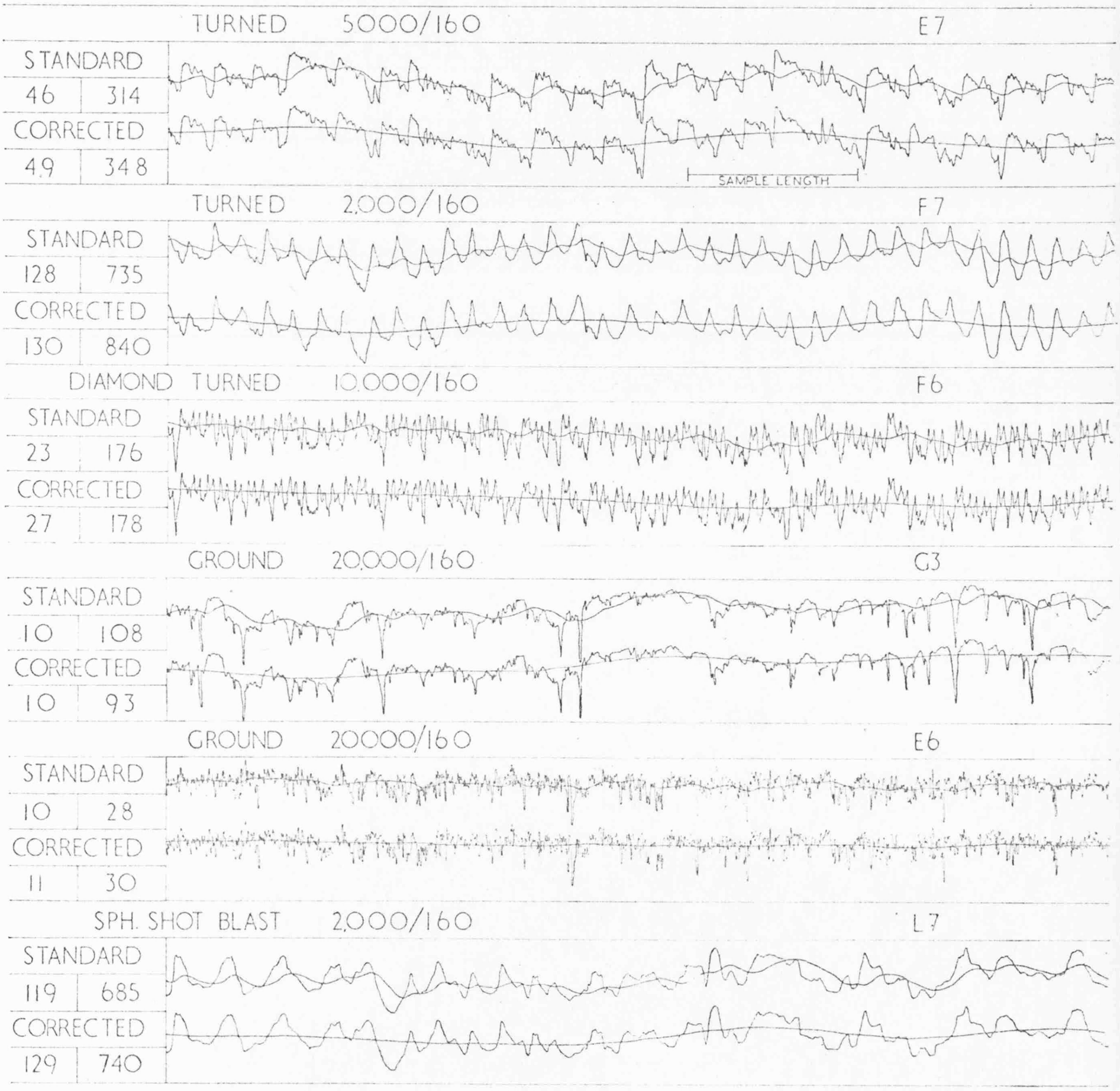


Fig. 12.8. Practical waveforms

The roughness which is transmitted by the phase-corrected filter actually looks like the roughness on the original profile and it is a sensible precaution to measure any desired parameter, even for waveforms containing components near to the cut-off. Another point is that the mean line for components within the cut-off is straight, with the result that all roughness within the cut-off is assessed. Yet another possibility offered by this filter is that the mean line of the roughness found by its use could properly be regarded as the waviness. This is because the shape of the mean line is only affected by components outside the cut-off (which are usually due to waviness) and also because the mean line will be in phase with these components. The mean lines for the repetitive waveforms outside the cut-off can be seen in Figs 12.6 and 12.7; Figs 12.8 and 12.9 show how the mean lines for practical profiles look convincing as the waviness.

On the basis of these properties it is suggested that the phase-corrected filter is particularly suited to the needs of surface measurement.

It has been shown how the phase-corrected wavefilter can assist in the interpretation and assessment of roughness. How can it be made? This question will be answered in the next section.

Practical models

Work has been carried out for a number of years by the Research Laboratories of the Rank Organization with a view to investigating and improving the behaviour of the

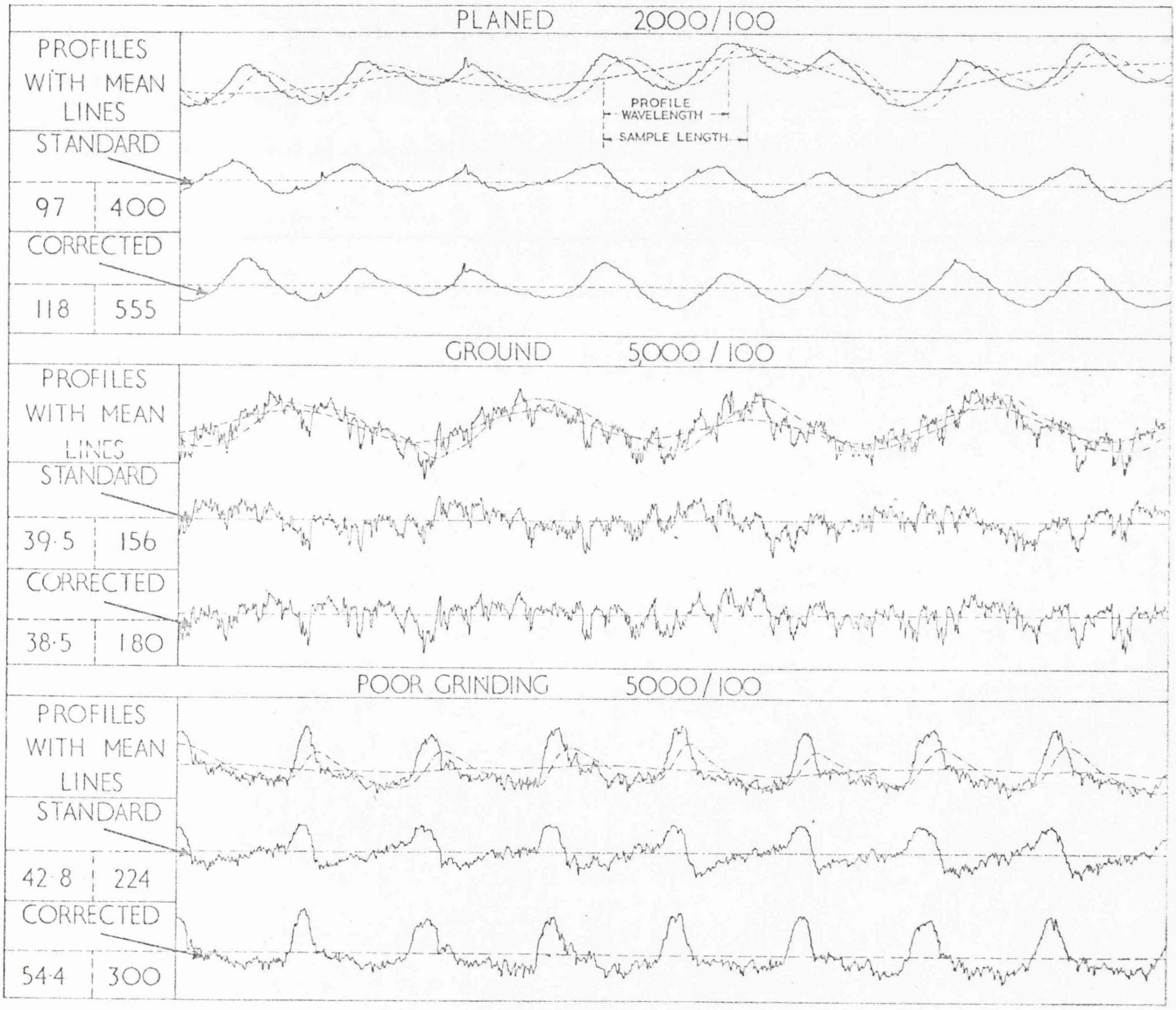


Fig. 12.9. Practical waveforms with components near to the cut-off

standard wavefilter mean line, especially the phase-distortion aspects. Attempts to improve the amplitude transmission characteristics using conventional minimum phase networks have usually resulted in worse phase distortion.

Eventually, a practical filter system having phase-corrected properties was conceived at the Rank Taylor Hobson Research Laboratory, and its effectiveness was established by means of computing techniques. The theory is described in Part 2. In this system the amplitude transmission characteristics could be adjusted at will without affecting the phase characteristics. Thus any characteristic could be investigated. Ways of realizing the filter were considered in conjunction with Rank Research Laboratories (Director Dr A. T. Starr), where two methods were subsequently developed and are now being investigated in a working model, the details of which must form the subject of a later publication.

Basically, the solution to the phase-distortion problem can be explained quite simply. For no phase distortion the constituent components of a waveform should not be shifted relative to each other in their passage through the filter. Obviously if all components passed instantaneously through the filter, then this criterion would be fulfilled. However, this is not possible in practice, the nearest possible achievement being to ensure that all components are delayed by the same amount. It can be shown (Part 2) that the criterion can be interpreted in another way. It is well known that the output from a filter depends not only on what is going in at any instant but also on what has been put in previously. Earlier signals put in at a time that is long in relation to the time constant of the filter have little or no effect on the filter's output, but events at earlier times which are small compared to the time constant of the filter have an effect. If the filter can be arranged so that the output at any instant is equally affected by future and past signals, then phase distortion does not take place.

It is interesting to note that the graphical procedure to determine the mean line called the mid-point locus mean line, devised by Reason as early as 1962, inherently incorporates the above principle and consequently did not suffer from phase distortion. Unfortunately the transmission characteristic was not suitable and sometimes led to an excessively wobbly mean line despite the phase correction (4).

As previously stated, the above requirement can only be achieved instrumentally by causing a uniform delay of the components of a profile in their passage through the filter. This is achieved by using a filter which has a phase-frequency characteristic that is linear. Filters of this type are called linear phase filters. Having this phase characteristic for the high pass filters used in surface-roughness measurement ensures that the mean line is in phase with the profile, and consequently has the advantages described.

As a basis for investigation, the amplitude transmission characteristics (Fig. 12.1*b*) mentioned earlier have been used. This has been judged useful from results found by computation of a number of surfaces. The transmission is unity up to the cut-off, then reduces at a rate proportional to the frequency up to three times the cut-off wavelength.

There are, of course, many other transmission characteristics that could be used. Some have already been investigated by computer, and it is possible that others will have to be considered, although the 3:1 attenuation characteristic has so far seemed quite acceptable.

The two practical embodiments which have now been built for comparative assessment will be called the time-reversal and transversal filter techniques. They have amplitude transmission characteristics to within 5 per cent of the desired transmission.

In the time-reversal technique the profile signal from the Talysurf is filtered using an ordinary wavefilter and is then recorded on magnetic tape. The tape is reversed in direction and the once-filtered signal is picked up and again passed through the filter. Phase distortion resulting during the first run of the signal through the filter is exactly cancelled out during the second because the signal is reversed in direction. The overall transmission of the signal is the square of the characteristic of the filter, which can be chosen to give any desired shape. In this arrangement the filtered signal finally emerging, backward, can be measured in the same way as an ordinary signal, but all parameters can now be measured from a realistic mean line.

The second method consists basically of using a low-pass filter and, in parallel, a system to enable the profile to be delayed. The former produces the mean line which is then subtracted from the delayed profile to give the filtered profile. In other words, an artificial time called 'now' that is slightly later than the real 'now' has been arranged for the filter. In this way the filter can know what is coming as well as what has gone—at least to the extent of the difference between the two 'nows'. Using this technique it is possible to fulfil the criterion for no phase distortion.

The transversal filter technique provides one method of synthesizing the low-pass filter. It is built up from a delay line made up from repetitive blocks of circuit elements. Any desired amplitude transmission characteristic or a close approximation to it can be realized.

Further work on these filters is in hand.

Subjects like this are more easily explained in mathematical terms than in words, and in Part 2 the theory of these phase-corrected or linear phase filters is shown together with some methods of making them by correlator and matched-filter techniques.

RESULTS

Surfaces which have roughness components unavoidably near to the cut-off are called Group I, while those which have their roughness well within the pass band are called Group II. Of about 50 surfaces examined, approximately twelve were in Group I. Fig. 12.9 shows three of these while Fig. 12.8 shows six other profiles, each with the mean lines found by the standard and phase-corrected filters. The number at the extreme left-hand side refers to the c.l.a. value and the other number refers to the maximum peak-to-valley height; both of these parameters are measured in micro-inches. The peak–valley parameter is used as an example of a parameter that is badly affected by phase distortion, although others could equally have been used.

To present some idea of how these parameters change, three types of surface have been considered. The first is for surfaces in Group I, the second for surfaces in Group II, and the third is for standard repetitive waveforms near to the cut-off. In this way, the two limiting cases will be covered. In Group II one would expect little difference between results from the two filters, whereas for the standard repetitive waveforms one would expect the most difference; the results from Group I should be somewhere between these two. As a measure of the effect of the distortion, the ratio of both the c.l.a. and peak–valley readings for the outputs of the standard and phase-corrected wavefilters were compared.

Taking first the case of the standard waveforms such as the sine, triangular, square, sawtooth, etc. it was found that the c.l.a. and peak–valley ratios are generally lower than unity, but not always. Both ratios could be as low as 0·5, while on the high side the peak–valley ratio could be 2·0. For the triangular waveform at the cut-off, for instance, the c.l.a. ratio is 0·82 and the peak–valley ratio is 0·5. The way in which these and other parameters are affected by distortion is not straightforward; for some waveforms the peak–valley value is accentuated by phase distortion (for the square wave), whereas for others it is reduced.

For surfaces of Group II the c.l.a. ratio was 0·98 with variations of about ±10 per cent and the peak-to-valley ratio 1·00±20 per cent, which means that well within the cut-off the results from the two filters are substantially the same. However, for those surfaces having roughness components near to the cut-off, the same average ratios were 0·85 and 0·75 respectively, showing that in these circumstances phase distortion does affect the measurement of the roughness. This can be seen quite clearly in Fig. 12.9*a*, for example, which consists of a planed surface having a periodic waveform near to the cut-off. The output from the standard filter has taken on a directional shape when compared to the original profile which is not realistic. Upon passing the same profile through the phase-corrected filter the roughness is completely undistorted and separated from the waviness.

A point to notice about Figs 12.8 and 12.9 is that the mean line found by the phase-corrected wavefilter seems to be a convincing basis for the measurement of the waviness.

CONCLUSIONS

The phase-corrected or linear phase wavefilter has the following advantages:

(1) The filtered profile within the pass-band is no longer distorted due to phase distortion, which means that the roughness signal emerging from the filter actually looks like the roughness on the profile. This helps interpretation considerably.

(2) Meaningful use of parameters such as bearing area and peak height in the region of the cut-off is now possible, which could be important in assessing some of the functional properties of the surface.

(3) Calibration of instruments is much more straightforward because any measurement of the filtered profile as indicated on a meter can agree closely with measurements taken on the chart with respect to a straight centre or mean line drawn within each sample length, and for repetitive waveforms can agree exactly for wavelengths right up to the cut-off.

(4) It is a plausible proposition to consider the mean line of the roughness as the waviness. The mean line of this filter only responds to undulations longer than the sample length and, further, it responds exactly in phase with these undulations. This means that the mean line immediately responds in a mechanically realistic way to any waviness present.

The behaviour of the phase-corrected filter to any waveform can be exactly calculated in a manner similar to the standard wavefilter (3), except that a different weighting function must be used.

The phase-corrected wavefilter represents an operative step forward in the evolution of the instrumentation of the M system. Unfortunately, it looks like being more costly, and the extent to which the extra realism introduced by its use will be worth the expense remains to be seen.

Part 2—Theory

INTRODUCTION

THIS PART BRIEFLY explains the theory behind the phase-corrected filter. It is approached from the time domain point of view because both the amplitude and phase characteristic can be investigated by manipulating the impulse response of the filter. Another reason for working in the time domain is that it enables a graphical procedure to be worked out which will enable the method to be applied to surface finish graphs. In this part the frequency equivalent of a wavelength on a surface is used because the impulse response transforms naturally into the frequency domain. Wavelength λ and frequency f can be converted by the simple formula

$$\lambda = v/f \quad . \quad . \quad . \quad . \quad (12.1)$$

where v is the scanning speed of the pick-up. Some of the properties of filters will be first discussed as an aid to the understanding of linear phase or phase-corrected filters.

Notation

B	Ratio of wavelengths having unity to zero transmission through the filter.
$C(\omega)$	Fourier transform of $c(t)$.
$c(t)$	Truncating function.
$F(\omega)$	Fourier transform of $f(t)$.
f	Frequency.
$f(t)$	Input to filter.
$G(\omega)$	Fourier transform of $g(t)$.
$g(t)$	Output from filter.
$H(\omega)$	Fourier transform of the impulse response.
$\bar{H}(\omega)$	Transfer function of high-pass filter.
$\underline{H}(\omega)$	Transfer function of low-pass filter.
$\bar{H}_c(\omega)$	Fourier transform of the truncated impulse response.
$\overset{*}{H}(\omega)$	Complex conjugate of $H(\omega)$.
$\bar{h}(t)$	Impulse response of high-pass filter.
$\underline{h}(t)$	Impulse response of low-pass filter.
$\bar{h}_c(t)$	Truncated impulse response.
$k(\omega)$	Amplitude transmission characteristic.
p	Laplace operator.
$R(\omega)$	Real component of transfer function.
t	Time variable.
t_0	Time delay.
v	Velocity of scan of pick-up.
$X(\omega)$	Imaginary component of transfer function.
α	Fraction of sample length.
$\bar{\alpha}$	Position of axis of symmetry.
α'	Variable under the convolution integral.
δ	Impulse of unit weight when integrated with respect to time.
δ'	Impulse of unit weight when integrated with respect to α.
λ	Wavelength
τ	Time variable under the convolution integral.
$\phi(\omega)$	Phase characteristic.
ω	Angular frequency variable.
ω'	Angular frequency variable under the convolution integral.
$*$	Convolution operation.
$\Leftrightarrow$	Indicates a transform pair.
$\mid\ \mid$	Process of taking the modulus.

FILTERS IN GENERAL

The frequency characteristics (or Fourier transform) $H(\omega)$ of a filter can be expressed as

$$H(\omega) = k(\omega)\,\mathrm{e}^{j\phi(\omega)} \quad . \quad . \quad . \quad (12.2)$$

where $k(\omega)$ is the amplitude transmission characteristic and $\phi(\omega)$ is the phase. The problem in surface metrology is to get a useful form for $k(\omega)$ while maintaining a form for $\phi(\omega)$ which eliminates phase distortion.

Impulse and step responses

These are the response of a filter to a unit impulse or step respectively. Either can be used to determine the output of a filter to any input. The output of the filter is given by the superposition integral, equation (12.3), which says in effect that the output at any time is given by the convolution of the input signal with the impulse response of the filter (5), i.e.

$$g(t) = \int_{-\infty}^{t} f(\tau)h(t-\tau)\,\mathrm{d}\tau$$

sometimes written

$$g(t) = f(t) * h(t) \quad . \quad . \quad . \quad (12.3)$$

where $f(t)$ is the input, $g(t)$ is the output, and $h(t)$ is the impulse response. τ is the time variable under the integral sign. Equation (12.3) is an example of time convolution; the similar operation, frequency convolution, is involved in the problem of impulse response truncation. In this equation $h(t-\tau)$, the reversed impulse response, is sometimes called the weighting function of the filter. It describes the amount of influence that a part of the signal has on the output of the filter at a time t. Equation (12.3) describes the property of filtering in the time domain.

If $H(\omega)$, $G(\omega)$, and $F(\omega)$ are the Fourier transforms of $h(t)$, $g(t)$, and $f(t)$ where

$$H(\omega) = \int_{-\infty}^{\infty} h(t)\,\mathrm{e}^{-j\omega t}\,\mathrm{d}t, \text{ etc.} \quad . \quad (12.4)$$

then

$$G(\omega) = F(\omega).H(\omega) \quad . \quad . \quad . \quad . \quad (12.5)$$

Another point about equation (12.3) is that it shows the similarity between the processes of correlation and filtering. The filtering process can be regarded either as the convolution of the input waveform with the impulse response of the filter, or the cross-correlation of the input waveform with the weighting function of the filter. The similarity is important where embodiments are concerned.

Equally, the profile could be considered to be made up of steps.

Equation (12.4) is very important because it shows that the frequency characteristics of a filter, i.e. its response to sinusoidal signals, are tied down to the shape of its impulse response. Altering the impulse response alters the frequency characteristics.

The impulse response can alternatively be expressed in terms of the fraction of the time constant of the filter which, in turn, can be expressed in terms of a fraction of the sample length when referred to the profile. This has the advantage of making not only the weighting function but also the variable in the superposition integral non-dimensional as well as relating more directly to the surface.

Different impulse responses and weighting functions

Consider a low-pass filter (Fig. 12.10*a*). It attenuates high frequencies. When referred to the time domain it means that the impulse response actually spreads in time. Because of this the impulse response tends to average out the high frequencies in the input waveform during the convolution process.

For high-pass filters (Fig. 12.10*c*) the situation is different because an impulse is present in the impulse response, which is opposed by a low-pass component. If the transfer function of an ordinary high-pass filter is $\bar{H}(p)$ it can be written in the form

$$\bar{H}(p) = 1-\underline{H}(p) \quad . \quad . \quad . \quad (12.6)$$

where $\underline{H}(p)$ is for a low-pass filter and p is the Laplace operator.

Equation (12.6) inverse transforms into an impulse response $\bar{h}(t)$, where

$$\bar{h}(t) = \delta-\underline{h}(t) \quad . \quad . \quad . \quad (12.7)$$

$\underline{h}(t)$ is the impulse response of the low-pass component and δ is an impulse at the origin of unit weight.

A signal $f(t)$ put into a high-pass filter gives an output

$$g(t) = \int_{-\infty}^{t} \bar{h}(t-\tau)f(\tau)\,\mathrm{d}\tau$$
$$= \int_{-\infty}^{t} \delta(t-\tau)f(\tau)\,\mathrm{d}\tau - \int_{-\infty}^{t} \underline{h}(t-\tau)f(\tau)\,\mathrm{d}\tau \quad (12.8)$$

but $\int_{-\infty}^{t} \delta(t-\tau)f(\tau)\,\mathrm{d}\tau = f(t)$—the sampling property of impulses. Hence equation (12.8) becomes

$$g(t) = f(t)-\int_{-\infty}^{t} \underline{h}(t-\tau)f(\tau)\,\mathrm{d}\tau = f(t)-m(t) \quad (12.9)$$

In practice the lower limit of the integral can be taken to be zero. Electrically $m(t)$ is the signal blocked by the filter. In surface metrology $m(t)$, when referred to the profile graph, is called the mean line. Removal of $m(t)$ from the input profile constitutes the high-pass filtering action, $\underline{h}(t-\tau)$ is the weighting function of the mean line (3).

Equation (12.9) can be expressed in a form more suitable for surface metrology by changing the time axis to a non-dimensional fraction of the sample length α:

$$g(\alpha) = \int_{0}^{\alpha} \delta'(\alpha-\alpha')f(\alpha')\,\mathrm{d}\alpha' - \int_{0}^{\alpha} \underline{h}(\alpha-\alpha')f(\alpha')\,\mathrm{d}\alpha'$$
$$= f(\alpha)-m(\alpha) \quad (12.10)$$

In equation (12.10) α' is a variable similar to τ in equation (12.9), i.e. $\bar{h}(\alpha) = \delta'-\underline{h}(\alpha)$ where δ' and $\underline{h}(\alpha)$ have unit weight when integrated with respect to α.

For the standard wavefilter

$$\bar{h}(t) = \delta-\frac{1}{RC}\left[2-\frac{t}{RC}\right]\mathrm{e}^{-t/RC} \quad . \quad (12.11)$$

or

$$\bar{h}(\alpha) = \delta'-A[2-A\alpha]\,\mathrm{e}^{-A\alpha} \quad . \quad . \quad (12.12)$$

where $A = \lambda/vRC$ (λ is the sample length, and v is the tracking speed), $\alpha = x/\lambda$ where x is the distance along the surface. In equation (12.11) both parts have the dimensions of reciprocal time whereas they are dimensionless in (12.12), which means that the ordinate scale does not change with the cut-off. The factor $1/T_c$ is taken into the variable $\mathrm{d}\alpha'$ of equation (12.10) where T_c is the equivalent time of the sample length.

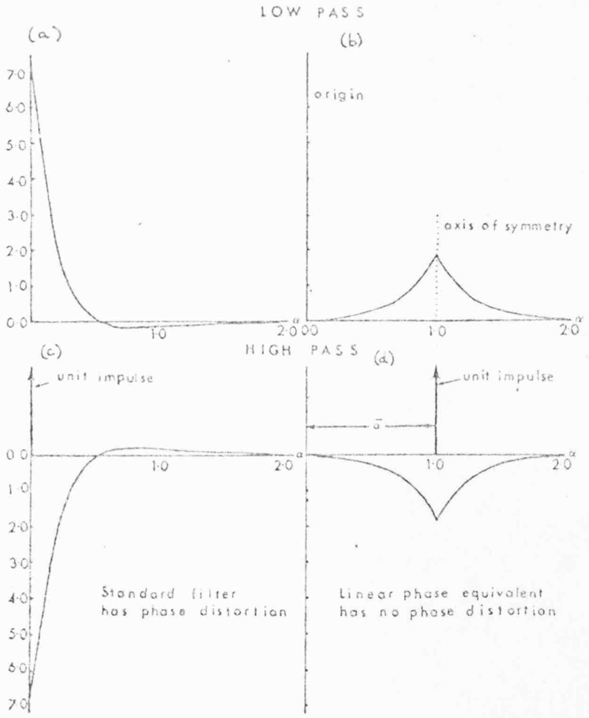


Fig. 12.10. Impulse response of linear phase and standard filter

LINEAR PHASE FILTERS (Fig. 12.10*b* and *d*)

The phase characteristics of a filter effectively show how sinusoidal signals get shifted in time in their passage through the filter. The real criterion for the filtered profile to be undistorted is that the constituent sinusoidal components making up the profile are not shifted relative to each other during the passage of the profile through the filter. One method of doing this would be to ensure that none of the components got shifted at all. This would mean that the phase shift is zero for all frequencies.

Now

$$H(\omega) = k(\omega)\,e^{j\phi(\omega)} = R(\omega)+jX(\omega) \quad (12.13)$$

where ϕ, the phase angle, is given by

$$\tan^{-1} X(\omega)/R(\omega) \quad . \quad . \quad . \quad (12.14)$$

So $X(\omega)$, the imaginary component, would have to be zero in equation (12.15) to make the phase zero, leaving

$$H(\omega) = k(\omega) = R(\omega) \quad . \quad . \quad (12.16)$$

But, for the transmission characteristic of the filter to be real, only the impulse response must be an even function, i.e. symmetrical about the time origin axis, which is impossible in practice because the impulse response or the weighting function of the filter cannot extend into the future as well as into the past. Hence, the only possibility of getting no distortion of the filtered profile is to arrange that the components of the profile are all shifted by the *same amount in time* by the filter. Suppose that this meant delaying all components by t_0. One component at angular frequency ω_1, say, would have to be shifted in phase by $-\omega_1 t_0$ rad to get this delay. A component of angular frequency ω_2 would have to be shifted by $-\omega_2 t_0$, and similarly for any component. In fact, to satisfy this delay of t_0 in general $\phi(\omega) = -\omega t_0$—the phase has to have a linear relationship with frequency. Therefore, the transmission characteristic for a filter having no phase distortion but a delay is of the form

$$H(\omega) = k(\omega)\,e^{-j\omega t_0} \quad . \quad . \quad (12.17)$$

Such a filter is called a linear phase filter.

How do the impulse responses of the zero delay and linear phase filters compare? It is easy to show that if $h_0(t)$ is the impulse response of the zero delay filter, then the impulse response of the linear phase filter having the same amplitude transmission is $h_0(t-t_0)$. This means that they have the same shape, but the linear-phase impulse response is shifted by t_0 along the positive time axis—they both have a symmetrical weighting function but the zero delay impulse response has its axis of symmetry on the time origin, whereas the linear phase impulse response has the axis of symmetry at $t = t_0$. Shifting the axis of symmetry to t_0 makes the impulse response practically realizable. Summarizing, it may be said that it is possible practically to make a filter giving no phase distortion only by allowing a uniform delay of all components passing through it. This is achieved by a filter having linear phase characteristics, which implies that it has an impulse response which is symmetrical about an axis of $t = t_0$ on the realizable side of the time axis. [In the non-dimensional form t_0 becomes $\bar{\alpha}$ (Fig. 12.10*d*).]

If $h_0(t) \Leftrightarrow H_0(\omega)$, then

$$h_0(t-t_0) \Leftrightarrow H_0(\omega)\,e^{-j\omega t_0} \quad . \quad (12.18)$$

For the zero delay case the high-pass impulse response $\bar{h}_0(|t|)$ can be expressed as $\delta - h_0(|t|)$, where $h_0(|t|)$ is the low-pass impulse response and δ is an impulse of the origin. The corresponding linear phase high-pass impulse response $\bar{h}_L(t)$ is therefore given by

$$\bar{h}_L(t) = \delta(t-t_0) - h_0(|t-t_0|) \quad . \quad (12.19)$$

Equation (12.9) becomes for a linear phase filter

$$g(t) = f(t-t_0) - \int_0^t \underline{h}_0(t-t_0-\tau) f(\tau)\,d\tau \quad (12.20)$$

Equation (12.19) shows that the impulse component lies at the axis of symmetry of the low-pass impulse component, which means that in terms of equation (12.20) the signal itself has to be delayed by t_0 before taking it from the low-pass component. This has the effect of producing at time t the filtered output—without distortion corresponding to the profile at $t-t_0$.

Two points are worth noting concerning symmetrical impulse responses: one is that the step response has an axis of odd symmetry about $t = t_0$, and the other is that the operations of correlation and convolution become the same except for an averaging term.

Different linear phase filters

The conditions for a symmetrical weighting function to be suitable are:

(*a*) It has an axis of symmetry later than $t = 0$ to an extent such that no considerable part of the function crosses the $t = 0$ axis.

(*b*) It is concentrated in a central lobe and dies away quickly on either side of the axis.

(*c*) Any negative portions should be small.

(*d*) It must have an amplitude transmission characteristic suitable for use in surface metrology.

A number of different linear phase filters have been investigated, including those which have Gaussian and raised cosine impulse responses (6) (7). Perhaps one of the most obvious is the linear phase filter having the same amplitude transmission as the standard filter. This was mentioned in Part 1, and Figs 12.3 and 12.6 show its mean line to two different profile waveforms.

It has an impulse response which by equation (12.19) becomes

$$\left.\begin{array}{l} \bar{h}(t) = \delta(t-t_0) - \dfrac{\omega_c}{\pi\sqrt{3}} \exp\left(-\dfrac{\omega_c}{\sqrt{3}}\,|t-t_0|\right) \\ \text{or alternatively} \\ \bar{h}(\alpha) = \delta'(\alpha-\bar{\alpha}) - \dfrac{\pi}{\sqrt{3}} \exp\left(-\dfrac{2\pi}{\sqrt{3}}\,|\alpha-\bar{\alpha}|\right) \end{array}\right\} \quad (12.21)$$

where the latter expression puts the impulse response in non-dimensional form, as in equation (12.3). ω_c is the angular frequency equivalent of the sample length, and $\bar{\alpha}$ is the position of the axis of symmetry—equivalent to t_0.

Another alternative is the ideal linear phase high-pass filter where

$$\left.\begin{aligned} \bar{h}(t) &= \delta(t-t_0)-\frac{\sin \omega_c(t-t_0)}{\pi(t-t_0)} \\ \text{or} \quad & \\ \bar{h}(\alpha) &= \delta'(\alpha-\bar{\alpha})-\frac{\sin 2\pi(\alpha-\bar{\alpha})}{\pi(\alpha-\bar{\alpha})} \end{aligned}\right\} \qquad (12.22)$$

These, however, did not seem to be as suitable a starting point as the following filter—called the phase-corrected filter in Part I. It has unity transmission for wavelengths up to the cut-off and zero transmission at three times the cut-off. The attenuation rate is proportional to the equivalent frequency (Fig. 12.1*b*).

The expression 'phase-corrected filter' has a slightly different connotation in communication theory, but it is useful in this context. The general equation for impulse responses having this form is

$$\left.\begin{aligned} \bar{h}(t) &= \delta(t-t_0)-\frac{2}{\pi\omega_c(1-B)} \\ &\quad \times\frac{\sin \omega_c(1+B)(t-t_0)/2\,.\sin \omega_c(1-B)(t-t_0)/2}{(t-t_0)^2} \\ \text{or} \quad & \\ \bar{h}(\alpha) &= \delta'(\alpha-\bar{\alpha})-\frac{1}{\pi^2(1-B)} \\ &\quad \times\frac{\sin \pi(1+B)(\alpha-\bar{\alpha}) \sin \pi(1-B)(\alpha-\bar{\alpha})}{(\alpha-\bar{\alpha})^2} \end{aligned}\right\} \quad . \quad . \quad . \quad (12.23)$$

where B is the ratio of wavelengths having unit to zero transmission, being equal to $\frac{1}{3}$ in this case. This factor has been found useful in surface roughness instruments to separate the roughness. As time goes on it may be that more useful characteristics emerge.

Realizable linear phase filters

To make a practical linear phase high-pass filter, an impulsive component and low-pass impulse response are usually involved. Consider first the impulse response of the low-pass component $R(t)$. To be realizable it must all lie to the right of the time axis $t = 0$. This is called the condition of causality. Hence the axis of symmetry t_0 (or $\bar{\alpha}$) should be delayed as long as possible. Also the response should decay quickly and not extend to infinity. Impulse responses of functions having transmission characteristics decaying faster than exponential do extend to infinity (8). Unfortunately, the theoretical impulse responses of those filters having the most suitable transmission characteristics do extend to infinity, so that even if the axis of symmetry is moved a long way from $t = 0$ the function is still finite at the $t = 0$ axis. Practically it has to end at $t = 0$, so that some degree of truncation of the impulse response from the ideal response must occur. Luckily, for the cases of interest, it is possible to get a useful approximation to the desired characteristic.

The effect of this truncation can be worked out using the technique of frequency convolution (5). It is the same problem as apodization in interferometry (9), and the determination of the power spectrum from the autocorrelation function (10). Truncation can be considered to be the multiplication of an ideal impulse function $h(t)$ by another function $c(t)$ called the truncation function of unit value for a limited region and usually zero elsewhere, which has the effect of modifying the ideal frequency characteristic $H(\omega)$. If $h(t)$ and $c(t)$ have Fourier transforms of $H(\omega)$ and $C(\omega)$, then the truncated impulse response $h_c(t)$ and the resultant frequency characteristic $H_C(\omega)$ are given by

$$h_c(t) = h(t).c(t) \quad . \quad . \quad . \quad (12.24)$$

and

$$H_C(\omega) = \int_{-\infty}^{\infty} H(\omega')C(\omega-\omega')\,\mathrm{d}\omega' \quad (12.25)$$

where ω' is a variable similar to τ in time convolution. Equations (12.24) and (12.25) are quite general. $H_C(\omega)$ is the approximation to $H(\omega)$ that results in practice. $c(t)$ need not be a simple box function, it can be any shape. The box truncation tends to produce oscillations in the transmission characteristics and also to reduce attenuation rates in $H(\omega)$, Fig. 12.11. Removal of this tendency can be achieved by a truncation function having no lobes in its Fourier transform. Unfortunately, the attenuation rates then become even more gradual. Truncation theory is important because it enables the minimum length of impulse response to be worked out, which allows the

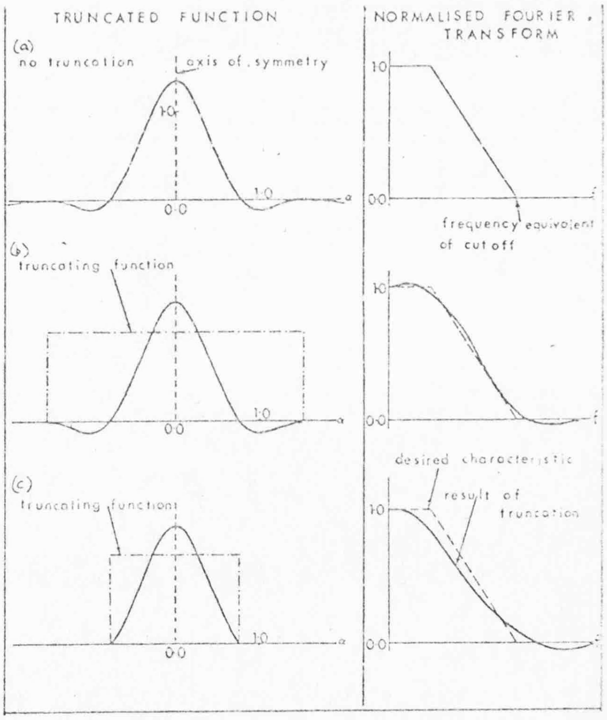


Fig. 12.11. Effect of truncation on low-pass component of phase-corrected filter

design criterion for the frequency characteristic to be achieved. In the case of the phase-corrected filter this is just three cut-off lengths to get the frequency characteristic $H_C(\omega)$ to within 5 per cent of $H(\omega)$.

The truncation problem only exists for the low-pass component, i.e. for the mean line evaluation. If the filtered profile is required, it is necessary to position the impulse component symmetrically relative to it. This means, in practice, that the signal has to be delayed by the amount by which the axis of symmetry of the mean line weighting function differs from $t = 0$. An exception is in the time-reversal case where the truncation of the impulse response depends on how long the tape can be run before assessment commences. The main approximation in this technique is made in the frequency domain where the square of the transmission characteristic of the filter has to fit as closely as possible to the desired characteristic.

Practical embodiments

There are many ways in which linear phase filters can be made. The methods used in making correlators and matched filters can be used because of the equivalence of convolution and correlation for symmetrical weighting functions. Some examples are in the use of optical, electrostatic, and electromagnetic storage devices with shaped masks, electrodes, and pick-up heads respectively. Magnetostrictive lines (11) or quantization correlators (12) could also be used. Digital techniques are an increasingly feasible proposition.

At the present time, however, electronic techniques are probably the best. Two techniques have been selected for assessment and a prototype model has been built incorporating both methods.

Time reversal. This reversal uses the principle that if a signal is passed through two filters, the one being the complex conjugate of the other, then there is no phase distortion. Here the signal is passed through the same filter twice but the conjugation is achieved by a reversal.

If a filter has an impulse response $h(t)$ whose transform is $H(\omega)$, then reversing the output can be shown to be equivalent to conjugation, with an additional phase linear term, i.e. if $h(t) \Leftrightarrow H(\omega)$, then

$$h(T-t) \Leftrightarrow e^{-j\omega T}\overset{*}{H}(\omega) \quad . \quad . \quad (12.26)$$

where $\overset{*}{H}(\omega)$ is the complex conjugate of $H(\omega)$. Hence passing the signal through a filter, reversing it, and then repassing it through the filter has the overall frequency characteristic of

$$e^{-j\omega T}\overset{*}{H}(\omega).H(\omega) = |H(\omega)|^2 e^{-j\omega T} \quad (12.27)$$

where T is the time taken to reverse the tape. Notice that this is still a linear phase filter because of the $e^{-j\omega T}$ term (13).

Transversal filter (14) In this the two components of the high-pass filter are constructed separately. The low-pass impulse response is synthesized in time by means of a delay line with a number of taps. Each tapping point is routed by way of an amplifier to a summing amplifier. Hence by adjusting these gains any impulse response can be built up. The impulse component is delayed to the centre of symmetry of the low-pass weighting function by means of a magnetic tape [equation (12.19)]. This just corresponds to delaying the signal [equation (12.20)]. The difference between the signal and the output from the summing amplifier is then taken to give the filtered profile. Because of dispersion, it is not usually possible to extract the signal from the position in the delay line which corresponds to the axis of symmetry of the weighting function and use it as the impulsive component.

ACKNOWLEDGEMENTS

It is a pleasure to give thanks to the Directors of Rank Taylor Hobson for permission to publish this paper. It is a pleasure also to acknowledge the contributions of Messrs Reason, Barr, and Spragg at Rank Taylor Hobson and of Dr Starr and Messrs Stillwell and Reeves at Rank Research Laboratories.

APPENDIX 12.1

REFERENCES

(1) British Standard 1134:1961. 'Assessment of surface texture'.

(2) American Standard B461:1962. 'Surface texture'.

(3) Whitehouse and Reason 'Equation of mean line of surface texture as found by an electric wavefilter', Rank Organization, 1965.

(4) Reason, R. E. 'Report of reference lines for roughness and roundness', *CIRP* 1962.

(5) Papoulis. *The Fourier integral and its applications* 1962 (McGraw-Hill).

(6) Sunde, E. D. 'Theoretical fundamentals of pulse transmission—I', *B.S.T.J.* 1964 (May).

(7) MacDiarmid, I. F. 'A testing pulse for television links', *Proc. Instn elect. Engrs* 1952.

(8) Lathi. *Signals, systems and communication* 1965 (Wiley and Sons).

(9) Mertz, L. (ed.). *Transformation in optics* 1965 (Wiley and Sons).

(10) Blackman and Tuckey. *The measurement of power spectra* 1958 (Dover).

(11) Monds and Rosie. 'Synthesis of matched filters for signal recognition using magnetostrictive delay lines', *Proc. Instn elect. Engrs* 1964 (No. 10).

(12) Allen and Westerfield. 'Digital compressed time correlators and matched filters for active sonar', *J.A.S.A.* 1964 36 (No. 1).

(13) Schreiner *et al.* 'Automatic distortion correction for efficient pulse transmission', *IBM Jl Res. Dev.* 1965 (January).

(14) Kallman, H. E. 'Transversal filters', *Proc. Inst. Radio Engrs* 1940 28, 302.

An investigation of the shape and dimensions of some diamond styli

J Jungles and D J Whitehouse
Rank Precision Industries Ltd, Metrology Research
Laboratory, Leicester House, Lee Circle, Leicester LEI 9JB

MS *received 20 November 1969, in revised form 5 March 1970*

Abstract An appraisal is given of the qualities of diamond styli which are used for surface texture measurement. Some of the difficulties encountered in normal optical and electron optical methods of assessment are discussed, and a possible solution which involves a two-stage replica process is described.

1 Introduction

The assessment of surfaces by tactile trace instruments has found a great deal of practical use in industry because of the ease of interpretation of profile graphs as well as the convenience of having an electrical signal available for the quantitative analysis of various parameters. The increasingly stringent demand for a better understanding of the relationships between surface conditions and the functional characteristics of component parts has been greatly aided by the developments during the last decade in electronics, which have served to make possible the design of more sophisticated instruments. We are now at a point in the development of profilometric measurement of ultra-fine surfaces where the dimensions in the contact region of the probe are of the utmost significance. The chart record which is revealed by the tactile tracer instrument can differ slightly from the true profile for a number of reasons, for instance the loading of the stylus can sometimes deform materials with low moduli of elasticity and plasticity. Another source of uncertainty is the lack of knowledge about the geometry of the stylus tip. For the normal run of engineering surfaces down to some 0·1 μm CLA, these effects have seemed negligible, but below this they become of increasing significance.

This paper describes more accurately the finite dimensions of the stylus tip. The aims in this endeavour have been twofold: firstly to perfect some techniques for the rapid assessment of the stylus tip and secondly to make use of these techniques in an attempt to set some kind of acceptance limit to which stylus tips can be worked by traditional methods.

The tips considered were (i) the nominal 2·5 μm Talysurf tip, and (ii) the ultra-sharp nominal 0·1 μm Talystep tip.

2 Methods of measurement of styli

2.1 *Traditional*

For the measurement of small geometrical forms using the light microscope it is often difficult to discriminate between real geometry and diffraction phenomena when objects have dimensions which are close to the theoretical limit of resolution. The method of light microscopy used for any specific application may reveal certain aspects of the object by phase or path differences. However, it is usually found that the more sophisticated forms of microscope will detract somewhat from the maximum possible resolution for a given wavelength. At each lens interface a certain amount of diffusion is produced by reflection, surface asperities and, in some cases, dust, so that the number of optical components should be kept to a minimum. For the particular problem of diamond tip assessment it has been our experience that those techniques which are contrived to increase the contrast of the image do so at the cost of resolution.

No matter which configuration of microscope is used for examining styli, it is important that the objective lens be of the best quality with a large numerical aperture. For establishing the quality of an objective lens we use the star test, in which a minute point of light is viewed against a dark background. A suitable test object is an evaporated film on a glass flat containing small pinholes.

Examination of one of the smaller points of light at each side of focus will reveal any spherical aberrations present in the objective lens. It is only when all the rays of light are focusing to a near point that the appearance of the diffraction effect on both sides of focus will be identical. The method of illumination for this test is not critical, apart from the need for a high intensity light source, because a small aperture in these circumstances can be considered as a self-luminous object. Our experience has been that an objective lens must perform well according to this test otherwise it cannot be exploited.

From an examination of the micrographs in figure 1, it is evident that bright field (*a*) does indeed show the major boundary best, while phase contrast (*b*) is overwhelmed by the relatively large dimensions of the pyramid faces. The Normarski interference method (*c*) (Lang 1968, 1969), whilst detracting slightly from the resolution of the boundaries, does show something of the texture on the tip itself. Figure 1(*d*) shows a chisel tip stylus as is used on the Talystep instrument for the measurement of very fine surface texture. This micrograph appears to indicate that the sharp edge of the chisel is about 0·2 μm across, but figure 3(*b*) reveals that it is less than 0·1 μm.

The extraction of reliable information from objects whose dimensions are close to the theoretical limits of the light microscope is difficult. When a craftsman is using a light microscope to monitor the various stages in the operations of diamond polishing, there is a practical limit of about 0·2 μm beyond which he may only proceed and estimate the degree of his success by guesswork.

2.2 *The scanning electron microscope*

For the analysis of the geometrical form of solid specimens the technique of electron bombardment of the object in order to generate secondary electron emission will have two main advantages over light microscopy: first, a large increase in the depth of focus is made possible by long-focus magnetic lenses which will keep the beam divergence small; second, an increase in the useful magnification will be attainable provided that the object is not too greatly affected by the electron beam, for example the object may acquire a charge or even suffer a change of its chemical composition.

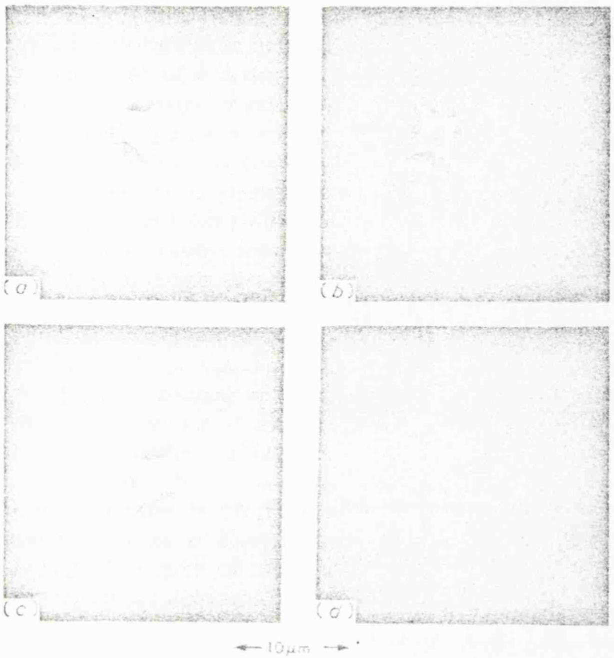


Figure 1 (*a*) Bright field, (*b*) phase contrast and (*c*) Normanski interference. For purposes of comparison, reference should be made to the electron micrograph (*a*) in figure 3 which is this same tip at a magnification of 9000×. (*d*) Chisel type stylus tip which can be compared with the electron micrograph (*b*) of this tip at 9000×. The objective lens used to obtain the micrographs in this figure was an exceptionally good 140X oil immersion of 1·3 numerical aperture

Among the variables which determine the resolution and contrast of the displayed image are shape, density, electrical conductivity and the secondary emission ratio of the various parts of the object, which is a function of the primary electron energy. Experimental data suggest that a primary electron at normal incidence to diamond and having an energy of 750 eV will release a maximum of about 2·8 secondary electrons. Ideally, what is wanted is a yield in energy which is unity with respect to the energy of a primary electron as this will minimize the tendency for potential gradients to develop. Secondary emission of an insulator such as diamond can be a complex phenomenon because of the absence of free electrons. An insulator will not lose energy as in a metal by interaction with free electrons in the conduction band. Primary electrons will only lose energy by interaction with valence electrons, and unless the insulating object is a thin film on an electrically conducting base, a space charge forms in the material. It is possible for an insulator to emit more electrons than were introduced, giving rise to a net loss of charge. The implications of this internal charging become evident when the geometry of the specimen is taken into account; charges will tend to collect or migrate towards peaks or other sharp boundaries.

The higher the energy of the primary (incident) electrons the further will they penetrate, and the number of electrons in the solid gaining energy from them will be greater. Electrons accelerated by 20 kV will penetrate solids such as aluminium to a depth of the order of 1 μm and it is expected that the penetration of electrons into diamond will be almost double this. There is evidence that surface contamination and thin conducting films used on specimens in scanning electron microscopes can greatly affect the emission properties; however, prevention of the charging effects in thick insulating objects would require an inordinate thickness of the conducting layers.

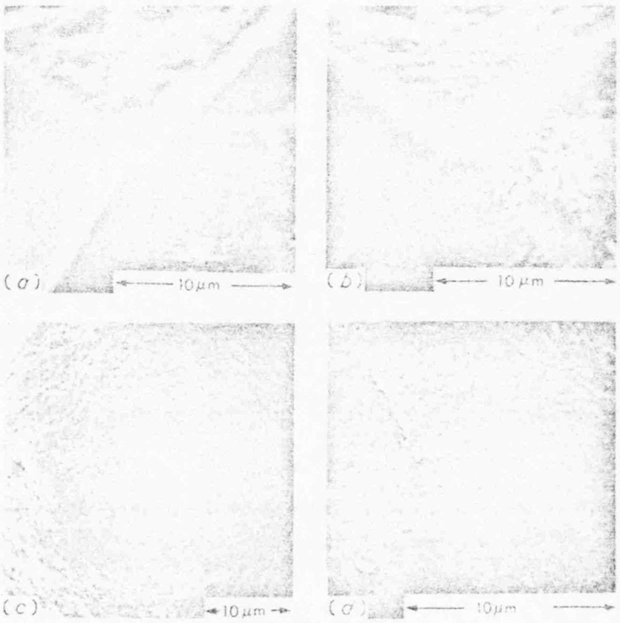


Figure 2 Electron micrographs from a stereoscan microscope: (*a*), (*b*) chisel tip stylus in the (*a*) uncoated and (*b*) the coated state. The coating is of aluminium and is about 20 nm thick; (*c*), (*d*) typical gramophone stylus whose tip radius is about 12 μm. The coating in this case is gold and about 10 nm thick

We find that overdeposited conductors such as aluminium, gold and carbon have little effect on the instrumental resolution at the diamond tip.

Figure 2(*a*) and (*b*) shows that the resolution at the extreme tip is somewhat increased by vacuum overcoating with 20 nm of aluminium; however, the gain is not sufficient for an accurate measurement of the sharp dimensions of the chisel to be made. The matter visible around the tip is dirt. We have sometimes found that an appreciable increase in resolution is evident in the near neighbourhood of such foreign substances.

The suitability of the scanning electron microscope for diamond tips on which there are no really prominent features is shown by plates (*c*) and (*d*), which illustrate a standard gramophone stylus having a tip radius of about 12 μm.

2.3 *The transmission electron microscope*

Some single-stage replicas were tried using the familiar materials, gelatine, collodion and Formvar, as well as some two-stage techniques which utilized wax, gelatine, acetate and metals as the first stage. Our lack of success with the various well-known techniques was for the most part due to the stringent requirement for a replica that not only showed the tip dimensions clearly but also a good deal of the pyramid leading to the tip. This sort of deep replica is easy to find in an area which may be littered with various shapes and markings.

Glass as a first-stage replica material was considered in view of the problems encountered with other materials. Examination of a range of glasses reveals that certain liquids become particularly viscous near their freezing point, thus preventing the formation and growth of crystal nuclei. If cooling of such a liquid is continued beyond a region where the mass becomes rigid, then the random atomic structure which is characteristic of glass will be held in the solid state. The amount of ordering of the structure of glass that is possible is strongly dependent upon the cooling rate, so that the final configuration of the glass will depend on its composition and thermal history.

The qualities of glasses are quite different from those of the organic polymers which are usually used for replicas in electron microscopy, these substances being partly crystalline and only to some degree amorphous. If a glass could be softened to flow around the tip of a diamond stylus, then a good replica might well be formed. The advantages of such a replica would be that an accurate representation of the tip should result because of the amorphous nature of glass and because of the ease of taking a second replica from the glass by the evaporated carbon technique.

As the hardness of glass (soda lime) is 530 on the Knoop scale, while diamond is placed at 7000 on this same scale, the relative hardness of diamond and glass will be about the same as that of tungsten carbide and zinc. Accordingly one might not expect damage to even a fine stylus tip when it is indented into glass, provided that no lateral shear forces can operate.

A small rig was constructed for use on a standard instrument (Talystep), which enabled a known load to be applied while the diamond was resting on the glass. The instrument could then be used to monitor the total amount of plastic deformation which occurred on application of load.

The criterion used to make good replicas is that the force is applied smoothly and within a short period of time. Having such control enables many such indentations to be made within a defined area. The replica is then shadowed and carbon is applied from at least two different directions to improve the rigidity of the replica. An example of how the replica can fracture on removal from the glass is shown in figure 3(*b*).

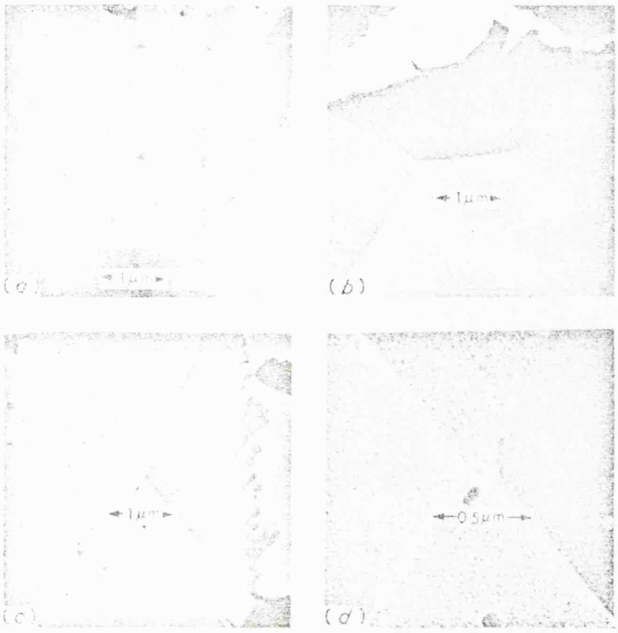


Figure 3 (*a*) Typical tip as used on our Talysurf instruments; (*b*) Talystep sharp stylus which is used for resolving very fine surface texture; (*c*) super-sharp stylus tip that was made in our laboratory; (*d*) an enlarged view of this very sharp tip

The stylus shown in figure 3(*a*) is the type of tip usually employed for surface texture measurement and is quite adequate for the assessment of surfaces produced by most machine processes used nowadays in industry. Figure 3(*b*) shows a stylus tip which is necessary to resolve the very fine texture associated with polished or lapped surfaces having a directional finish. In figure 3(*c*) and (*d*), a stylus of exceedingly small tip dimensions is shown. With such a tip it is possible to measure very fine randomly orientated surface asperities such as are found on evaporated films and in certain conditions of devitrification which take place during annealing processes, given the proviso that only very low stylus forces can be used.

3 Discussion

Use of the more precise information about styli obtained by the methods described has already made possible a better understanding of some of the smaller features on surfaces (Whitehouse and Archard 1970).

We now know with a fair degree of certainty, just how far the spatial resolving power of the tactile tracer instrument can be effectively taken. Under normal loading conditions this limit will approach 0·1 µm or about $\frac{1}{4}$ of the wavelength of ultraviolet light. Further information can be obtained by virtue of the fact that if the stylus dimensions are known then they can be partly computed out of a profile chart. The argument does however have limitations in that on the one hand the positive asperities can have the stylus error removed but negative valleys into which the stylus cannot fully penetrate must be computed, although such computation can be made reasonably accurately.

The effects of two different stylus tip dimensions can be seen in figure 4 where chart (*a*) shows the profile of a small section of gauge block as measured by the relatively large tip (figure 3(*a*)), and chart (*b*) is a portion of this same surface as measured by the fine chisel tip stylus (figure 3(*b*)). Comparison of the charts (*a*) and (*b*) clearly shows that the

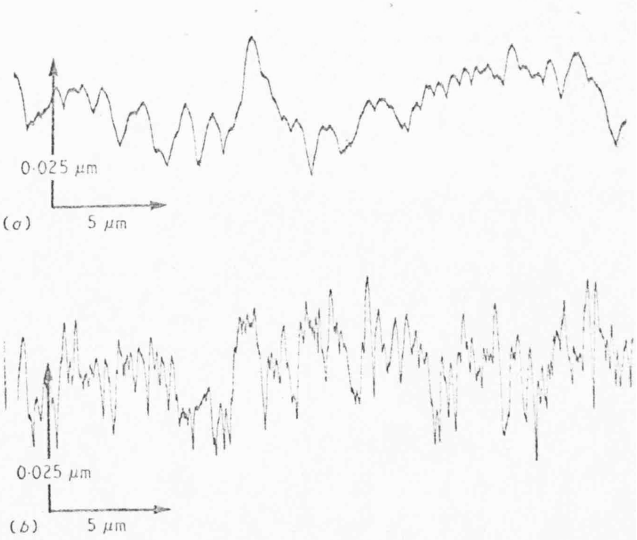


Figure 4 (*a*) Profile of a small section on a finely lapped gauge block using the stylus shown in figure 3(*a*); (*b*) a section of this same surface as measured by the stylus shown in figure 3(*b*) larger stylus tip rides over and loses fine detail. A knowledge of tip dimensions is desirable in microhardness testing, where within the range of 100 g to 1 mg loads, indents are made into specific inclusions within polycrystalline bodies.

In ultramicrotomy where diamond knives are used to section specimens for electron microscopy, this method might play a part in defining the quality of the diamond knives used and their rate of wear.

Acknowledgments

Our thanks to Dr Arthur Rowe and Mrs Doris Duncan of Leicester University for giving valuable advice and help in obtaining the electron micrographs, to Dr Reason of our Research Laboratory who first suggested trying glass and Mr R C Spragg for his help and encouragement. Finally our thanks also to the Directors of Rank Precision Industries for permission to publish this paper.

References

Lang W 1968 *Zeiss Information* **70** 114

Lang W 1969 *Zeiss Information* **71** 12

Whitehouse D J and Archard J F 1970 *Proc. R. Soc., Lond.* **A316** 97–121

REFERENCES

ABBOTT, E.J. and FIRESTONE, F.A., Specifying Surface Quality. Mech.Engng. (1933).

AL-SALIHI, T., Ph.D. Thesis, University of Birmingham (1967).

AMERICAN STANDARD, ASA B46.5

AMONTONS. Histoire de l'Academie Royal des Sciences avec les mémoires de mathématique et de physique. p.206 (1699).

ARCHARD, J.F., Proc.Roy.Soc. A 243 190 (1957).

ARCHARD, J.F., The crossed-cylinders friction machine. Wear 2, 21. (1958).

BAUL, R.M., Mechanics of metal grinding with particular reference to Monte Carlo simulation. 8th Int.M.T.D.R. (1967).

BECKMANN, P., A new approach to the problem of reflection from a rough surface. Acta Technica No. 4 p.311 (1957).

BECKMANN, P., The probability distribution of the vector sum of n unit vectors with arbitrary phase distributions. Acta Technica No. 4 p.323 (1959).

BECKMANN, P. and SPIZZICHINO, A., The scattering of electromagnetic waves from rough surfaces. Pergamon (1963).

BENDAT, J.S., Principles and applications of random noise theory. p.213. Wiley (1958).

BENDAT, J.S. and PIERSOL, A.G., Measurement and analysis of random data. Wiley (1966).

BENNETT, H.E. and PORTEUS, J.O., J.Opt.Soc.Am. 51, 123 (1961).

BLACKMAN, R.B. and TUKEY, J.W., The measurement of power spectra. Dover New York (1958).

BLOK, H., Proc.Roy.Soc. Lond. A.212 p.480 (1952).

BOWDEN, F.P. and TABOR, D., The friction and lubrication of solids. Part 1. Oxford University Press (1954).

BUTLER, R.D. and POPE, R.J., Surface roughness and lubrication in sheet steel metalworking. Proc.Inst.Mech.Engrs. 182 Pt. 3K 162 (1968).

BUTTERY, T.C., Grinding and abrasive wear. Ph.D. Thesis. Dept. of Engineering, University of Leicester. (1969).

CRAMER, H., Mathematical methods of statistics. Princeton University Press (1946).

COULOMB. Memoires de Mathematique et de Physique de l'Academie Royale des Sciences. p.161 (1785).

DAVENPORT, W.B. and ROOT, W.L., An introduction to the theory of random signals and noise. McGraw Hill (1958).

DAVIES, H. Proc. Inst.Elec.Engrs. 101, 209 (1954).

EHRENREICH, M. The slope of the bearing area as a measure of surface texture. Microtecnic (1959).

GREEN, A.P., Proc.Roy.Soc. A.228 191 (1955)

GREENWOOD, J.A. and WILLIAMSON, J.B.P., Contact of nominally flat surfaces. Proc.Roy.Soc.Lond. A.295 p.300 (1966).

GREENWOOD, J.A. and TRIPP, J.H., The elastic contact of rough spheres. J.App.Med. p.153 (March 1967).

GRIEVE, D.J., KALISZER, H. and ROWE, G.W., A "normal wear" process examined by measurement of surface topography. Annals C.I.R.P., Vol.XVII (1969).

GUMBEL, D.J., Statistics of extremes. Columbia University Press (1959).

HALIDAY, J.S., Proc.Inst.Mech.Engrs. 169 p.177 (1955).

HAYKIN, S.S. and CARNEGIE, R., A new method of synthesising linear digital filters based on the convolution integral. Proc.I.E.E. Vol.117 No.6 p.1063 (1970).

H.M.S.O., Interpolation and allied tables. (1956).

HOLM, R., Electric Contacts Handbook. Springer-Verlag (1958).

HOWELLS, R.I.L. and PROBERT, S.D., Contact between solids. T.R.G. Report 1701 (R/X). (1968)

IWAKI, A. and MORI, M., On the distribution of surface roughness when two surfaces are pressed together. Bull.Japan S.M.E. Vol.1 No.4 p.329 (1958).

JUNGLES, J. and WHITEHOUSE, D.J., An investigation of the shape and dimensions of some diamond styli. Jo. of Physics.E. Scientific Instruments. Vol.3 p.440 (1970).

KIMURA, Y., An analysis of the distribution of contact points through the use of surface profiles. J.Japan S.Lub.Eng. 11(11) p.467. (1966).

KRAGELSKII, I.V., Friction and wear. Butterworths (1965).

KRENDEL, E.S., I.R.E. Trans.Med.Elect. p.149 (1959).

KUBO, M., Rev.Sci.Inst. 36, 236 (1965).

KUBO, M. and PEKLENIK, J., An analysis of microgeometric isotropy for random surface structures. CIRP Ann. Vol.16 p.235 (1968).

LEE, Y.W., Statistical theory of communication. Wiley (1960).

LINDEN, D.A., A discussion of sampling theorems. Proc.I.R.E. 1219 (1959).

LINNIK, Y. and KHUSU, A.P., Mathematico-statistical description of surface profile irregularity in grinding. Inzhernernyi. Sborn. 20, 154 (1954).

LONGUET-HIGGINS, M.S., Statistical analysis of a random moving surface. Proc.Roy.Soc.Lond. Vol.249A 966 p.321 (1957).

LIU, S.C., Bell System Tech.Journal. Dec. p.2273 (1968).

LUNN, M.S., Some experiments on the relation between friction and surface finish. Project Report (1969). Eng.Dept. University of Leicester.

MARTIN, L.C., Theory of the microscope. Blackie (1967).

McADAMS, H.T., PICCIANO, L.A. and REESE, P.A., A computer method for hypsometric analysis of abraded surfaces. Proc. 9th M.T.D.R. Conf. p.73 Pergamon (1968).

MERCHANT, M.E., The mechanism of static friction. J.App.Phys. 11, 230 (1940).

MYERS, N.O., Characterisation of surface roughness. Wear 5 182 (1962).

NAGASU, H., Statistical features in static friction. Bull.Jap.S.Mech.Engrs. Vol.6 (1951).

NARA, J., Some analysis of paper finished surfaces. J.S.P.M.J. 28, 120 (1962).

OSTVIK, R. and CHRISTENSEN, H., Changes in surface topography with running-in. Paper 8 p.59. Proc.Symposium on Tribology. Trondheim (1968).

PANTER, P.F., Modulation, noise and spectral analysis applied to information transmission. McGraw Hill (1965).

PAPOULIS. Probability, random variables and stochastic processes. McGraw Hill (1965).

PARZEN, E., Stochastic processes. Holden-day Inc. (1962).

PEKLENIK, J., Contribution to the theory of surface characterisation. CIRP Ann. 12, 173 (1963).

PEKLENIK, J., Investigation of the surface typology. CIRP Ann. 15, 381 (1967).

PEKLENIK, J. and KUBO, M., A basic study of a 3-D assessment of the surface generated in a manufacturing surface. CIRP Ann. Vol.16 p.257 (1968).

PESANTE, M., Determination of surface roughness typology by means of amplitude density curves. CIRP Ann. 12,61 (1963).

RABINOWICZ, E., The nature of the static and kinetic coefficients of friction. J.App.Phys.Vol.22 No.11 (Nov.1951)

RABINOWICZ, E., RIGHTMIRE, B.G., TEDHOLM, C.E., and WILLIAMS, R.E., The statistical nature of friction. Paper 54 ASME/ASLE (1954).

RABINOWICZ, E., Autocorrelation of the sliding process. J.App.Phys.Vol.27. No. 2 (Feb. 1956).

RABINOWICZ, E., Direction of the frictional force. p.1073 Nature No. 4569 (May 1957).

REASON, R.E., GARROD, R.I., and HOPKINS, M.R., A report on the measurement of surface finish by stylus methods. Rank Precision Industries Ltd. (1944).

REASON, R.E., The Measurement of Surface Texture. Modern Workshop Technology. MacMillan (1969).

REASON, R.E., Report on reference lines for roughness and roundness. CIRP Ann BDXI (1962).

REASON, R.E., (a) Le Calcul automatique des critères des profils de surfaces. Automatisme 9 No. 5. 177 (1964).
(b) The bearing parameters of surface topography. Proc.5th M.T.D.R. Conf. (1964).

REASON, R.E., Modern Workshop Technology. p.622. MacMillan (1969).

RENEAU, J. and COLLINSON, J.A., B.S.T.J. Vol.XLIV No.10. p.2203 (1965).

RICE, S.O., The mathematical analysis of random noise. B.S.T.J. 23 282-332 (1944).

RICE, S.O., The mathematical analysis of random noise. B.S.T.J. 24 46-156 (1945).

SHARMAN, H.B., Calibration of surface texture measuring instruments. Proc.I.Mech.Eng. Vol.182 Pt 3K(1968).

SHARMAN, H.B., Influence of sample size and the relationships between the common surface texture parameters. Proc.I.Mech.E. Vol.182 Pt. 3K (1968).

SPRAGG, R.C., and WHITEHOUSE, D.J., A new unified approach to surface metrology. To be published by I.Mech.E. (1971).

TOLANSKY, S., Multiple-beam interference microscopy of metals. Academic Press (1970).

TSUKIZOE, T. and HISAKADO, T., On the mechanism of contact between metal surfaces. Trans.A.S.M.E. p.666 (Sept.1965).

WATTS, D.G., A general theory of amplitude quantisation with applications to correlation determination. I.E.E. Monograph 481M p.209 (1961).

VON WEINGRABER, H., Suitability of the Envelope Line as a standard for measuring roughness. Microtecnic 1 (1957).

VON WEINGRABER, H., Accuracy and reliability of roughness measurements. CIRP Ann. Vol.18 (1969).

WHITEHOUSE, D.J., Typology of manufactured surfaces. CIRP Vol.18 (1970).

WHITEHOUSE, D.J., An improved wavefilter for the measurement of surface texture. Proc.I.Mech.E. Vol.182 Pt 3K. (1968).

WHITEHOUSE, D.J. and REASON, R.E., Equation of the mean line of surface texture found by an electric wave filter. Rank Precision Industries Ltd. (1965).

WIDROW, B., A study of rough amplitude quantisation by means of Nyquist Sampling Theory. I.R.E. Trans.Circuit Theory (Dec.1956).

WILLIAMSON, J.B.P., Topography of solid surfaces. Interdisciplinary approach to friction and wear. NASA Symposium SP-181 (Nov. 1967).

WILLIAMSON, J.B.P., Microtopography of surfaces. Proc.I.Mech.E. Vol.182 Pt 3K (1968).

WILLIAMSON, J.B.P., PULLEN, J. and HUNT, R.T., The shape of solid surfaces. Proc.A.S.M.E. Los Angeles (Nov. 1969).

WIRTZ, A., Ein Beitrag Zur Typologie der Oberfläche. CIRP Ann. 17, 307 (1969).

WORMERSLEY, J.R. and HOPKINS, M.R., J.Etats Surface 135 (1945).

YOSHIKAWA, H. and SATA, T., Simulated grinding process by Monte Carlo method. CIRP Ann. Vol.16 p.297 (1968).

YOSHIKAWA, H. and PEKLENIK, J., Three-dimensional simulation techniques of the grinding process. Pt.I. Presented to CIRP General Assembly, Nottingham 1968.

YOSHIKAWA, H. and PEKLENIK, J., Three-dimensional simulation techniques of the grinding process. Pt.II. CIRP Ann. Vol.18 (1970).