Contact of nominally flat surfaces

By J. A. Greenwood and J. B. P. Williamson

Burndy Corporation Research Division, Norwalk, Connecticut, U.S.A.

(*Communicated by F. P. Bowden, F.R.S.—Received* 28 *April* 1966)

It is usually assumed that the real area of contact between two nominally flat metal surfaces is determined by the plastic deformation of their highest asperities. This leads at once to the result that the real area of contact is directly proportional to the load and independent of the apparent area—a result with many applications in the theories of electric contacts and friction. Archard pointed out that plastic deformation could not be the universal rule, and introduced a model which showed that, contrary to earlier ideas, the area of contact could be proportional to the load even with purely elastic contact.

This paper describes a new theory of elastic contact, which is more closely related to real surfaces than earlier theories. We show how the contact deformation depends on the topography of the surface, and establish the criterion for distinguishing surfaces which touch elastically from those which touch plastically. The theory also indicates the existence of an 'elastic contact hardness', a composite quantity depending on the elastic properties and the topography, which plays the same role in elastic contact as the conventional hardness does in plastic contact.

A new instrument for measuring surface topography has been built; with it the various parameters shown by the theory to govern surface contact can be measured experimentally. The typical radii of surface asperities have been measured. They were found, surprisingly, to be orders of magnitude larger than the heights of the asperities. More generally we have been able to study the distributions of asperity heights and of other surface features for a variety of surfaces prepared by standard techniques. Using these data we find that contact between surfaces is frequently plastic, as usually assumed, but that surfaces which touch elastically are by no means uncommon in engineering practice.

Introduction

It has long been realized that surfaces are rough on a microscopic scale, and that this causes the real area of contact to be extremely small compared to the nominal area. The calculation of the area of contact, or even the prediction of how this varies with load, is very difficult. Early attempts to do this by applying the Hertzian theory of contact between spheres to individual contact spots met with two difficulties: the area of the contact spot depends on the radius of the asperity, which is not usually known; and the predicted variation of area with load proved to be incorrect. Both these obstacles were removed when Holm introduced the idea that although the overall stresses are in the elastic range the local stresses at the contact spots are much higher so that the elastic limit will be exceeded and the contact will yield plastically; each contact can then be visualized as a small hardness indentation, so that the mean contact pressure will be equal to the hardness and effectively independent of the load and the contact geometry. This has been a highly fruitful concept, both in the study of electric contact and, as developed by Bowden & Tabor (1954, 1964), in the related subject of friction.

It was in this related study that the objection to the concept was first put forward. Archard (1957) pointed out that although it is reasonable to assume plastic flow for the first few traversals of one body over another it is absurd to assume this for machine parts which may make millions of traversals during their life: the asperities may flow plastically at first, but they must reach a steady state in which the load is supported elastically. He went on to show that although the simple Hertzian theory did not predict the observed proportionality between contact area A and load P, a generalized model in which each asperity is covered with microasperities, and each microasperity with micromicroasperities, gave successively closer approximations to the law $A \propto P$ as more stages were considered. Archard explained that the essential part of the argument was not the choice of asperity model: it was whether an increase in load creates new contact areas or increases the size of existing ones; for physically plausible surfaces any elastic model in which the number of contacts remains constant will give $A \propto P^{\frac{2}{3}}$: but if the average size remains constant (and the number increases) the area will be proportional to the load.

Thus contact between flat surfaces can be determined either by plastic or by elastic conditions; and we may expect that for very rough surfaces there will certainly be plastic flow, while for very smooth ones, contact will be entirely elastic. This paper presents a more detailed model of elastic contact between nominally flat surfaces, and uses it to determine where the changeover from elastic to plastic contact occurs. Nominally flat surfaces may be defined as those in which the area of apparent contact is large so that the individual contacts are dispersed and the forces acting through neighbouring spots do not influence each other. When curved surfaces touch the apparent contact area is limited by their gross geometry; the individual areas are then tightly clustered and it becomes necessary to take account of the interaction between them (Greenwood & Tripp, in the press).

Mathematical model

We shall consider the contact between a plane and a nominally flat surface covered with a large number of asperities which, at least near their summits, are spherical. We assume that all asperity summits have the same radius β, and that their heights vary randomly: the probability that a particular asperity has a height between z and $z+\mathrm{d}z$ above some reference plane will be $\phi(z)\,\mathrm{d}z$. Figure 1 shows schematically the type of contact envisaged. The behaviour of an individual asperity

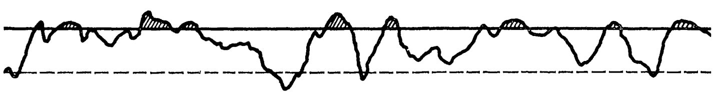


FIGURE 1. Contact of rough surfaces. The load is supported by those asperities (shaded) whose heights are greater than the separation between the reference planes.

is known from the Hertzian equations (Timoshenko & Goodier 1951). The contact radius a_1, area A_1, and load P_1 can be expressed in terms of the compliance w (the distance which points outside the deforming zone move together during the deformation) as

$$a_1 = \beta^{\frac{1}{2}} w^{\frac{1}{2}}, \quad A_1 = \pi \beta w, \quad P_1 = \tfrac{4}{3} E' \beta^{\frac{1}{2}} w^{\frac{3}{2}}$$

where

$$\frac{1}{E'} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}.$$

If either of the contacting surfaces is much more elastic than the other, E' is just the 'plane-stress modulus' for that material, $E/(1-\nu^2)$: if the materials are the same E' is half this.

If the two surfaces come together until their reference planes are separated by a distance d, then there will be contact at any asperity whose height was originally greater than d. Thus, the probability of making contact at any given asperity, of height z, is

$$\text{prob}\,(z > d) = \int_d^\infty \phi(z)\,\mathrm{d}z$$

and if there are N asperities in all, the expected number of contacts will be

$$n = N\int_d^\infty \phi(z)\,\mathrm{d}z.$$

Also, since $w = z-d$ and $A_1 = \pi\beta w$, then the mean contact area is

$$\int_d^\infty \pi\beta(z-d)\,\phi(z)\,\mathrm{d}z,$$

and the expected total area of contact will be given by

$$A = \pi N\beta\int_d^\infty (z-d)\,\phi(z)\,\mathrm{d}z.$$

Similarly, we find the expected total load is

$$P = \tfrac{4}{3}NE'\beta^{\frac{1}{2}}\int_d^\infty (z-d)^{\frac{3}{2}}\,\phi(z)\,\mathrm{d}z,$$

and if we assume electrical contact over the whole of the area of mechanical contact, then since the conductance of a single contact is $G_1 = 2a/\rho$ where ρ is the resistivity (see Holm 1958) the expected total conductance will be

$$G = 2N\rho^{-1}\beta^{\frac{1}{2}}\int_d^\infty (z-d)^{\frac{1}{2}}\,\phi(z)\,\mathrm{d}z.$$

Since this argument assumes that the microcontacts are sufficiently separated to be mechanically independent it seems reasonable to treat the current flow through them also as independent.

It is convenient to introduce standardized variables, and describe heights in terms of the standard deviation σ of the height distribution, also, we introduce the surface density of asperities η and write $N = \eta\mathscr{A}$ where $\mathscr{A}$ is the nominal contact area. This gives

$$\begin{aligned} &\text{number of contact spots} && n = \eta\mathscr{A}F_0(h);\\ &\text{total conductance} && G = 2\eta\mathscr{A}\rho^{-1}\beta^{\frac{1}{2}}\sigma^{\frac{1}{2}}F_{\frac{1}{2}}(h);\\ &\text{total contact area} && A = \pi\eta\mathscr{A}\beta\sigma F_1(h);\\ &\text{load} && P = \tfrac{4}{3}\eta\mathscr{A}E'\beta^{\frac{1}{2}}\sigma^{\frac{3}{2}}F_{\frac{3}{2}}(h); \end{aligned}$$

where h, the standardized separation, is equal to d/σ and

$$F_n(h) = \int_h^\infty (s-h)^n\,\phi^*(s)\,\mathrm{d}s,$$

where $\phi^*(s)$, the standardized height distribution, is the height distribution scaled to make its standard deviation unity.

Exponential distribution of asperity heights

One interesting case (which provides a convenient example in which the above equations can be treated exactly) is when the heights follow an exponential distribution: then $\phi^*(s) = \mathrm{e}^{-s}$. The functions $F_n(h)$ are just $n!\,\mathrm{e}^{-h}$, and so we have

$$n = \eta\mathscr{A}\,\mathrm{e}^{-h}, \qquad G = \pi^{\frac{1}{2}}\eta\rho^{-1}(\beta\sigma)^{\frac{1}{2}}\mathscr{A}\,\mathrm{e}^{-h},$$
$$A = \pi(\eta\beta\sigma)\mathscr{A}\,\mathrm{e}^{-h}, \quad P = \pi^{\frac{1}{2}}(\eta\beta\sigma)E'(\sigma/\beta)^{\frac{1}{2}}\mathscr{A}\,\mathrm{e}^{-h}.$$

Eliminating the separation h we find that there is exact proportionality between the load and the number of contact spots, the conductance, and the area of contact. Thus, the average size of contact spots, and the contact pressure, are independent of the load. Of course, the size of any individual contact spot increases with load but at the same time new, small, spots are formed and there is a perfect balance which leaves the average unchanged.

This result does not depend on the particular surface model or deformation mode considered: it holds provided the asperities have an exponential height distribution and all obey the same area/compliance and load/compliance laws. Although it will be shown later that height distributions tend to be Gaussian rather than exponential, the exponential distribution is nevertheless a fair approximation to the uppermost 25 % of the asperities of most surfaces. This leads us to suggest that the origin of the laws of friction, and particularly of the proportionality between area and load, lies not in the ideal plastic flow of individual contact spots but simply in the statistics of surface roughness. This speculation is discussed more fully by Greenwood (1965)

Gaussian distribution of asperity heights

The experimental results which follow show that for many surfaces the height distribution is Gaussian to a very good approximation. We have, therefore,

$$\phi^*(s) = \frac{1}{\sqrt{(2\pi)}}\,\mathrm{e}^{-\frac{1}{2}s^2},$$

and so

$$F_n(h) = \frac{1}{\sqrt{(2\pi)}}\int_h^\infty (s-h)^n\,\mathrm{e}^{-\frac{1}{2}s^2}\,\mathrm{d}s.$$

The functions $F_{\frac{1}{2}}(h)$ and $F_{\frac{3}{2}}(h)$ had not been tabulated, and so were computed by integration of the differential equation satisfied by $\exp(-\frac{1}{2}h^2)\,F_n(h)$, namely

$$y'' - hy' - (n+1)\,y(h) = 0;$$

we worked backwards from the values at large h which may be found from the asymptotic expansions. Subsequently tables of parabolic cylinder functions appeared (Miller 1964) and were used to check the values.

From these equations, and using physically reasonable values for the parameters, we have calculated the relations between the load and the separation, the area of contact, the mean pressure, and the contact resistance. The results approximate

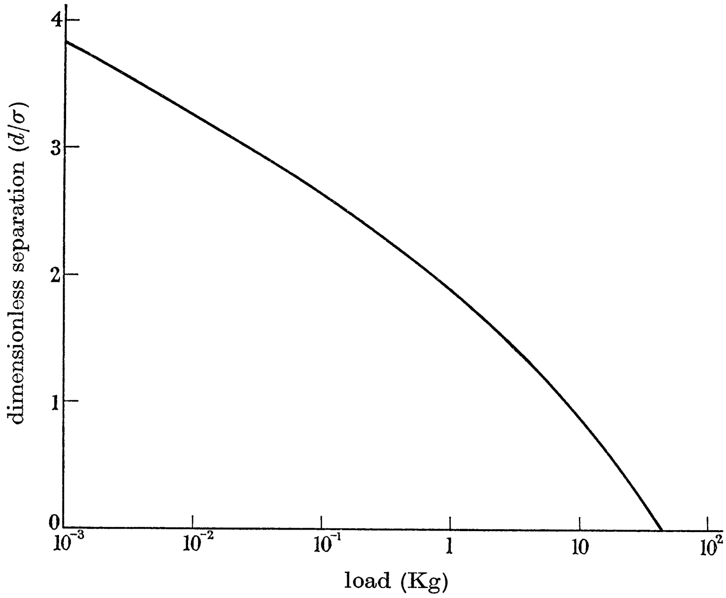


FIGURE 2 (*a*). Relation between separation and load. The curve shows the expected value of the load for a given separation, assuming a Gaussian distribution of asperity heights.

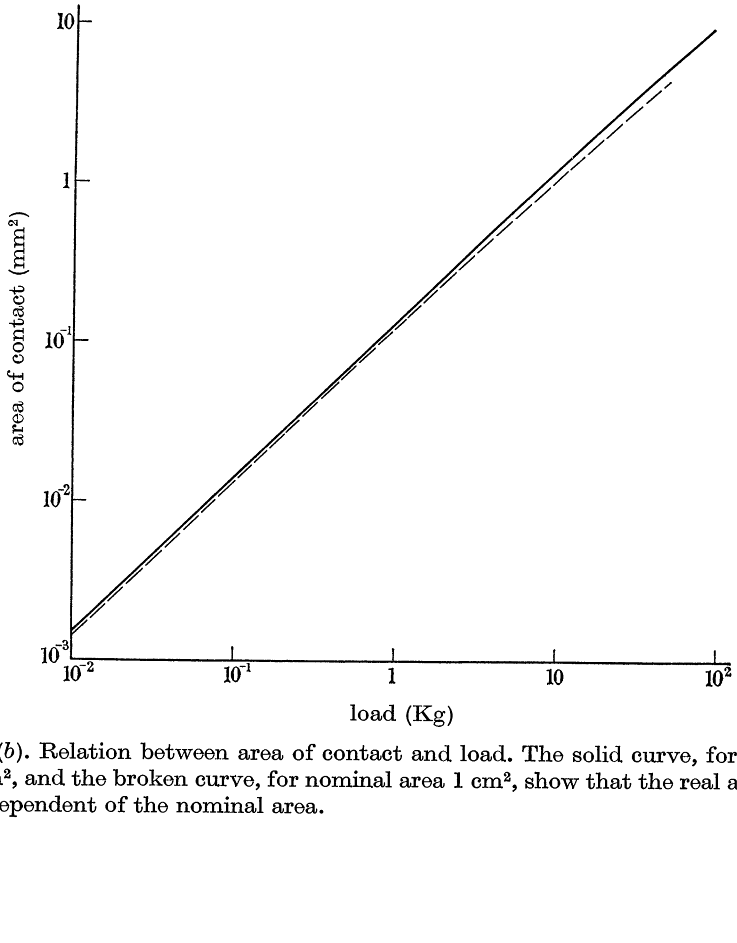


FIGURE 2 (*b*). Relation between area of contact and load. The solid curve, for nominal area 10 cm^2, and the broken curve, for nominal area 1 cm^2, show that the real area of contact is independent of the nominal area.

closely to those for the exponential distribution; they are shown in figures 2 to 4, for the case

$$\eta = 300/\text{mm}^2, \quad \beta\sigma = 10^{-4}\,\text{mm}^2,$$
$$E'(\sigma/\beta)^{\frac{1}{2}} = 25\,\text{Kg/mm}^2, \quad \rho = 2{\cdot}4\,\mu\Omega\text{ cm}.$$

Figures 2(*b*), 3, and 4 each show the relation between the expected values of two variables both dependent on the separation. To have calculated the expected value of one for a given expected value of the other would have been very much more difficult: preliminary work using a Monte Carlo method suggests that except at low loads the two approaches give the same result.

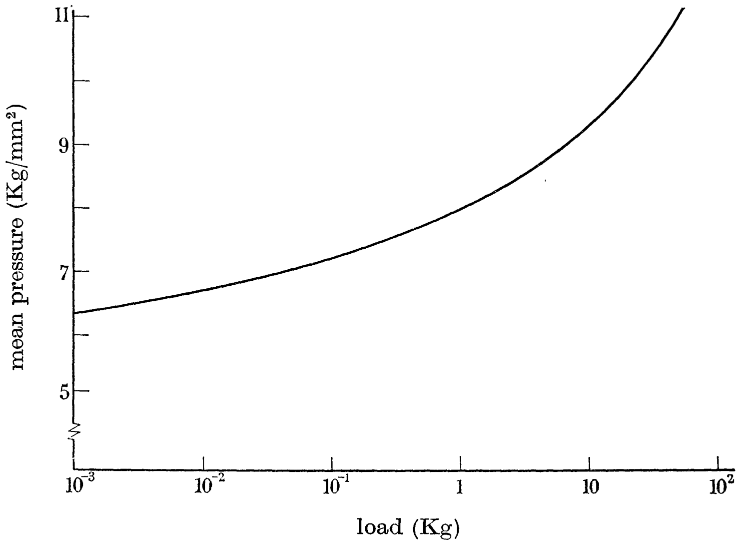
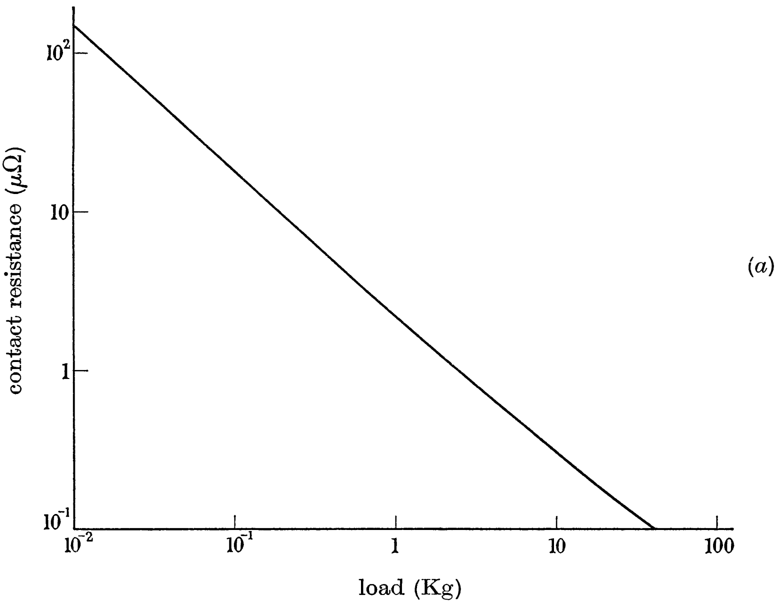


FIGURE 3. Variation of mean real pressure with load. A change of the load by a factor of 10^5 causes the real pressure to change by only a factor of 2.

Figure 2(*a*) shows how the separation varies with the load, calculated for a nominal area of $1\,\text{cm}^2$: an increase from a light pressure of $100\,\text{g/cm}^2$ to a heavy pressure of $10\,\text{Kg/cm}^2$ merely reduces the separation from $2{\cdot}6\sigma$ to $0{\cdot}9\sigma$. Figure 2(*b*) shows that for a given nominal area the area of contact is almost exactly proportional to the load. Further, the two curves, for nominal contact areas of 1 and $10\,\text{cm}^2$, are almost indistinguishable, showing that the contact area depends on the load and not on the nominal pressure. The similarity with the behaviour calculated by assuming ideal plastic flow, where the mean pressure is equal to the hardness, suggests that there is an 'elastic contact hardness' which controls the area of contact under elastic conditions. The validity of this concept is indicated in figure 3, which shows how the mean pressure at the contact areas, P/A, varies with load. For a load range of 10^5 the mean pressure varies only between $0{\cdot}2$ and $0{\cdot}4$ of $E'(\sigma/\beta)^{\frac{1}{2}}$. The elastic contact hardness, at least for a Gaussian surface, can be taken as $0{\cdot}25E'(\sigma/\beta)^{\frac{1}{2}}$; it plays the same role in relations describing the elastic contact of rough surfaces as does the hardness in discussions of plastic contact. Using this concept, we can calculate the area of contact directly from the load for elastic contact precisely as



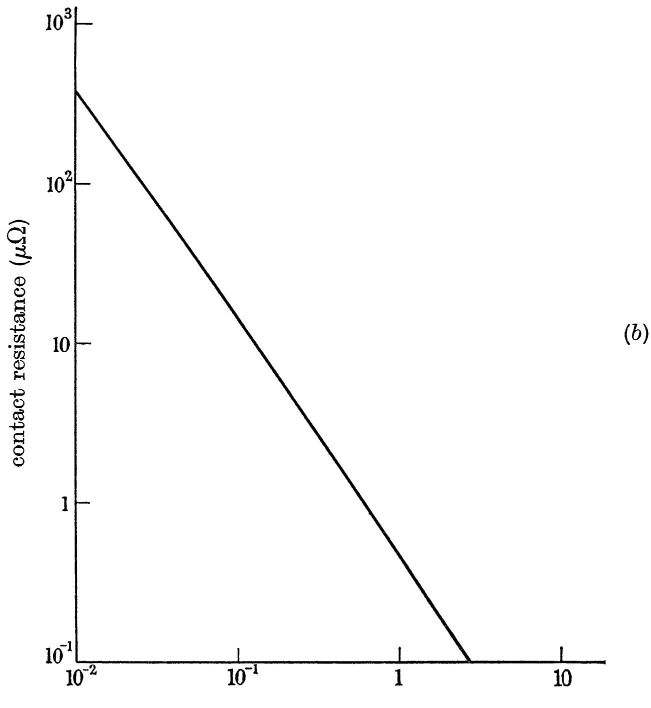


FIGURE 4. Variation of contact resistance with load. Figure 4 (*a*) shows the resistance when the current flows through all mechanical contact spots, as in contact between film-free surfaces; this gives $R \propto P^{-0.9}$. Figure 4 (*b*) shows the result of assuming that only the plastically deformed contacts pass current, as may happen with film-covered surfaces; this gives $R \propto P^{-1.4}$.

we should for plastic contact using the hardness. Figure 4*a* shows the contact resistance, R, assuming that the mechanical contact areas are all conducting. The result approximates closely to the simple law $R \propto P^{-0 \cdot 9}$.

The separation depends on the nominal pressure and not on the total load. The variables n, G, and A, on the other hand, depend on both the nominal area and the separation, and in such a way that their overall dependence on the nominal area is small. Thus, the separation is the link which, while depending itself on the nominal pressure, enables the other variables to depend on the load only. In rough terms the dependence on pressure can be explained by assuming that a certain number of contact spots is needed to provide a given contact area: when the nominal area is large there will be enough high peaks to provide these, while for a small nominal area lower peaks will be needed also.

Contact of rough curved bodies

When the contact areas are limited to a small part of the surface because of the overall curvature of the bodies, the assumption that the individual contacts are independent fails: the contacts will be sufficientlyclose together for the force on one to change the height of its neighbours. The treatment is then much more involved than that for independent contacts (Greenwood & Tripp, in the press). It shows that the classical Hertzian contact theory is a high load limit for rough surfaces, and that at lower loads the area of contact is much more dispersed. The *real* pressures at the microcontacts (which are much higher than the Hertzian pressures) are of the same order as those for nominally flat surfaces; and are again almost proportional to $E' \surd(\sigma/\beta)$. In both cases the factor varies slowly with load: for the contact between curved surfaces it also varies with the roughness; but for all the cases considered the range was only 0·36 to 0·61, in good agreement with the values 0·2 to 0·4 found for nominally flat surfaces.

The origin of the slow variation of the contact pressure is the same as it is for nominally flat surfaces: again, the number of contacts is roughly proportional to the load and the average size of a contact area is almost constant. Thus the arguments in this paper are not limited to the nominally flat surfaces considered.

Limit of elastic deformation

From the work of Tabor (1951) on the ball indentation hardness test, we know that the onset of plastic flow is reached when the maximum Hertzian pressure q_0 between a ball and a plane reaches about $0 \cdot 6H$, where H is the hardness. Since

$$w = (\tfrac{1}{2}\pi)^2 q_0^2 \beta / E'^2$$

the critical value of the elastic displacement at the asperity necessary for some plastic flow is

$$w_p = 0 \cdot 89 \beta (H/E')^2.$$

Plastic flow first occurs internally, and will be restricted by the surrounding elastic material, so we may conveniently take the criterion for detectable plastic flow to be rather higher, i.e. at

$$w_p = \beta (H/E')^2.$$

Now, just as the probability of making contact at a given asperity is

$$\text{prob}\,(z > d) = \int_d^\infty \phi(z)\,\mathrm{d}z,$$

so the probability of a plastic contact is

$$\text{prob}\,(z > d+w_p) = \int_{d+w_p}^\infty \phi(z)\,\mathrm{d}z.$$

Again, the expected total area of the contacts which become plastic will be

$$A_p = \pi\eta\beta\mathscr{A}\int_{d+w_p}^\infty (z-d)\,\phi(z)\,\mathrm{d}z$$

$$= \pi\eta\beta\sigma\mathscr{A}\int_{h+w_p^*}^\infty (s-h)\,\phi^*(s)\,\mathrm{d}s$$

where $$w_p^* = w_p/\sigma = (\beta/\sigma)\,(H/E')^2.$$

Provided that only a small proportion of the area of contact is plastic, the overall properties will be close to those predicted by the elastic theory. Moreover, the predictions concerning the plastic contacts will be reasonably good, because the plastic displacements will be limited to an elastic magnitude, since any asperity which starts to collapse will immediately have its excess load taken by the remaining asperities. Accordingly, we can use elastic theory to study the growth of plastic contact areas with increasing load. For example, for the surface used for figures 2 to 5, assuming a hardness such that $w_p^* = 2$, we find that 1 % of the contact area is plastic at a nominal pressure of 1·1 Kg/cm², 2 % at a pressure of 3·7 Kg/cm², and 5 % at a pressure of 13·4 Kg/cm². For $w_p^* = 1{\cdot}5$, the transition from elastic to plastic contact is even slower, the corresponding pressures being 18, 122, and 1200 g/cm².

If we define the limit of elastic contact to be when the area of plastic contact becomes some specified fraction of the total contact area, we can obtain a relation between the surface roughness, as defined by the factor w_p^*, and the critical value of the nominal pressure. The fundamental relation is in fact between the surface roughness and the critical separation h_c, for the ratio A_p/A depends only on w_p^* and h, with no other parameters entering. Thus, we arbitrarily choose a value of A_p/A, determine the critical separation as a function of w_p^*, and hence, introducing the values of the surface parameters, we obtain the corresponding nominal pressure.

The factor w_p^* is slightly unsatisfactory as a generalized surface roughness parameter, since it decreases when the roughness increases, and so we shall substitute $\psi = (w_p^*)^{-\frac{1}{2}} = (E'/H)\sqrt{(\sigma/\beta)}$ which we call the plasticity index: it combines the material and topographic properties of the solids in contact. Table 1 (*a*) shows how the critical separation varies with the plasticity index of the surface taking $A_p/A = 0{\cdot}02$ as the criterion for the onset of a significant degree of plasticity. The nominal pressure which would have to be applied to cause the necessary critical separation is also shown.

The critical value of the plasticity index for a given nominal pressure varies with the particular value of A_p/A used as the criterion, but as table 1 (*b*) shows, the variation is not important.

In principle the plasticity index merely determines the critical load at which the deformation changes from elastic to plastic. In practice, however, the plasticity index completely dominates the behaviour, and the load has little effect; for, although the plasticity index can in theory have any value from 0 to ∞ (and it appears that surfaces range from 0·1 to over 100) it is only for the narrow range 0·6 to 1 that the mode of deformation is in doubt. When ψ is less than 0·6 plastic contact could be caused only if the surfaces were forced together under very large nominal pressures indeed: when ψ exceeds 1·0 plastic flow will occur even at trivial nominal pressures. Most surfaces have plasticity indices larger than 1·0, and thus, except for

TABLE 1. LIMIT OF ELASTIC DEFORMATION OF ROUGH SURFACES

(*a*) *Variation of separation h_c and critical nominal pressure $p_{\text{nom.}}$ with plasticity index ψ*

ψ	0·6	0·7	0·8	0·9	1·0
$h_c = d/\sigma$	0·11	1·34	2·57	3·80	5·08
$p_{\text{nom.}}$(g/cm^2)	$4·30 \times 10^4$	$3·75 \times 10^3$	$1·22 \times 10^2$	1·07	$2·0 \times 10^{-3}$

For $\psi < 0·6$ the pressures are impracticably high, and for $\psi > 1·0$ they are impracticably low.

(*b*) *Variation of plasticity index ψ with elastic breakdown criterion A_p/A for $p_{\text{nom.}} = 1\,\text{Kg/cm}^2$*

A_p/A	0·01	0·02	0·05	0·10	0·50
ψ	0·70	0·74	0·81	0·88	(1·26)

The value for $A_p/A = 0·5$ is in parentheses since it is unreasonable to use elastic theory so far beyond the elastic limit.

(All values of the nominal pressure have been calculated assuming $\eta E'\beta^{\frac{1}{2}}\sigma^{\frac{3}{2}} = 0·75$, in agreement with the values of η, $\beta\sigma$, and $E'(\sigma/\beta)^{\frac{1}{2}}$ used in figures 2 to5.)

especially smooth surfaces, the asperities will flow plastically under the lightest loads, as has frequently been postulated. These values, and indeed the definition of the plasticity index, hold only for the particular surface model considered, but it seems clear that the concept is a general one.

Plastic deformation may be of vital importance even when the area of plastic contact is trivial. Frequently with oxide covered surfaces electrical contact will occur only at plastically deformed asperities. Thus the conductance through the plastic contacts, G_p, is of great interest. This is given by

$$G_p = 2\eta\mathscr{A}\rho^{-1}(\sigma\beta)^{\frac{1}{2}}\int_{h+w_p^*}^{\infty}(s-h)^{\frac{1}{2}}\phi^*(s)\,\mathrm{d}s$$

Figure 4 (*b*) shows how the contact resistance varies with the load, for the same case as figure 4 (*a*) with added condition that $\psi = 0·7$. The graph shows the conductance through the areas of plastic contact under conditions where the contact as a whole is still within the limits of validity of the elastic theory, i.e. $A_p/A < 0·02$. The resistance is a thousand times larger than the value for oxide-free surfaces, and it varies rather more rapidly with load: $R \propto P^{-1·4}$. This appears to be the first time an exponent less than minus one has been derived theoretically, though such behaviour, over limited ranges, is well known experimentally.

In real surfaces there is a range of asperity radii. The effect on the purely elastic properties is very small, since the influences of greater and smaller radii largely cancel; but the effect on the plastic properties is more serious: the sharper asperities may flow plastically when the ones with the mean radius do not, and so there can be no averaging. This may reduce the dependence of the resistance on load.

There is an interesting formal similarity between the plasticity index suggested in this paper and the usual criterion for elastic surface contact due to Blok (1952). Blok considers a model surface of equal parallel sinusoidal ridges, and shows that these can be completely levelled elastically provided the average slope is less than $H/\pi E'$. Blok's yield condition, that the maximum contact pressure equals the hardness, H, reduces to the condition that the maximum (internal) shear stress is equal to the yield stress in shear, k, provided the hardness equals $2ek \sim 5{\cdot}4k$, which is a good approximation (see Tabor 1951). Expressing the average slope in terms of the ridge height, h, and the radius of curvature at the summit, β, Blok's rule for purely elastic contact can be written

$$\frac{E'}{H}\sqrt{\frac{h}{\beta}} < 1.$$

Halliday (1955) has developed a similar 'average slope' criterion for a single spherical asperity, again based on the requirement of complete elastic flattening, and by applying this to individual asperities observed by reflexion electron microscopy has demonstrated the existence of elastic surfaces. There is, however, a fundamental difference between these approaches and the present theory. The condition that asperities must be completely elastic is too stringent. Purely elastic contact can occur between surfaces where the asperities could become plastic if deformed separately. The present approach shows how the complete criterion com bines both the shape of individual asperities and the spread of their heights.

Measurement of surface topography

The approach developed above indicates that the contact behaviour of a surface can be described in terms of two material properties: the hardness and the elastic modulus, and two topographic parameters: the radius of asperity summits and the spread of asperity heights. The usual way of specifying surface texture is to give a single parameter, such as the mean deviation from an average line. This is, of course, quite inadequate for the application of surface contact theory. In principle the information required is available in a profilometric trace (Abbott & Firestone 1933) such as that shown in figure 5. It is possible to derive the distribution of peak heights manually from traces such as these, and to find the mean radius of the peaks by manual curve-fitting; but the process is tedious, and it would be impracticable to generate data in useful quantities.

To overcome this difficulty we have built a surface analysing system, which basically consists of a Taylor–Hobson model 3 Talysurf feeding a digital computer through a suitable analog-to-digital conversion and sampling unit. The voltage analog of the surface is obtained in the usual way by means of a stylus and an electromechanical transducer. A graph of the profile is plotted by a pen recorder, and at the same time the data conversion unit samples this voltage and punches it on paper tape. The sampling rate is 36·7 readings per second, each reading being three decimal

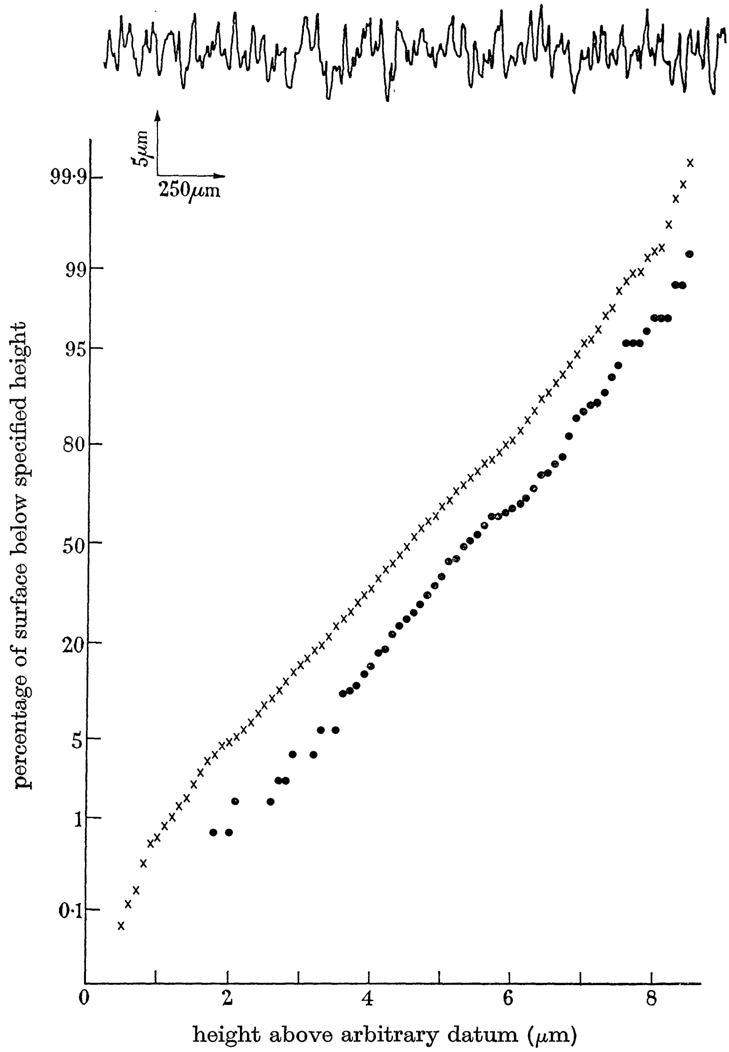


FIGURE 5. Cumulative height distribution of bead-blasted aluminium. Both the distributions of all heights (×) and of peak heights (●) are Gaussian, at least in the range ± 2 standard deviations. The profile of the same surface is shown in the upper diagram: the vertical magnification is 50 times the horizontal magnification.

digits. With the usual horizontal magnification of 100 this records the surface height every 1·7 μm: with the 500 times magnification the horizontal resolution becomes 0·34 μm. Autocorrelation measurements show high correlation between adjacent height readings, which indicates that the horizontal resolution is adequate. We have not yet found a surface for which the full resolution of 0·34 μm was needed. In a typical observation 1500 height readings are recorded from a profile corresponding to 0·1 in. of the surface. At the largest vertical magnification the height is recorded in units of 10Å. Stringent precautions have to be taken to isolate the stylus from mechanical vibrations, and the Talysurf transducer is mounted on an elaborate aseismic table. In practice the system is limited by residual noise—mechanical and electrical—with an amplitude of about 50Å. Similar systems are now in use at Messrs Rank Taylor Hobson, and at the National Engineering Laboratory (R. E. Reason; B. Sharman, private communications).

The computer is programmed to evaluate many different surface texture parameters: for this analysis it locates the peaks in the profiles, and calculates σ the standard deviation of their heights, η their surface density, and β their mean radius; for comparison it also calculates the height distribution of the entire surface, and the conventional 'centre line average'.

Experimental study of surface topography

A series of experiments was performed to determine the height distribution, the peak height distribution, and the plasticity indices of representative surfaces; and also to study how ψ changes during various surface treatments. Topographic data obtained from several typical surfaces are given in figures 5 to 7. Figure 5 shows the surface profile and the height distributions for a bead-blasted aluminium specimen (c.l.a. 47 μin.). The heights are measured from an arbitrary reference plane, and the proportion of the surface lying below a specified height is plotted on normal probability paper, which has a distorted scale so designed that a Gaussian distribution of heights will appear as a straight line. The height distribution of the surface (crosses) is indeed Gaussian. This agrees with results reported by Bickel (1963). The circles show the distribution of the peaks: this too is Gaussian. The standard deviation, σ, is 1·37 μm; the mean radius of the peaks is 13 μm. Applying the above theory we find that the surface has a plasticity index of 30, and would thus deform plastically at all loads.

Although several common surface preparations produce Gaussian distributions, many do not; figure 6 shows the results for a mild steel specimen which had been slid against a copper flat under oleic acid. Neither the surface nor the peaks are Gaussian, although even in this case the uppermost peaks form a reasonably good approximation to a Gaussian distribution. In fact it appears to be generally true that the peak height distribution of a surface is more nearly Gaussian than is the overall surface height. Although the surface is not particularly smooth (3·4 μin. c.l.a. σ is 0·065 μm) the peak radius is so large (0·24 mm) that the plasticity index is only 0·8.

Surfaces with values of ψ well into the elastic range occur commonly: for example, a one inch diameter steel roller bearing was found to have a c.l.a. of 1·6 μin., a standard deviation of 0·024 μm, and a mean peak radius of 150 μm; this gives a plasticity index of 0·25. The very low value of ψ was in this case partly due to the great hardness. However, softer materials can also be highly elastic. A polished mild steel specimen with a c.l.a. of 0·5 μin. showed a mean peak radius of 0·5 mm, a standard deviation of 100Å, and a plasticity index of 0·3. Holm long ago reported mean pressures of 0·02 H for carefully polished surfaces (see Holm 1958); since for elastic contact the mean pressure is about $0{\cdot}25E'(\sigma/\beta)^{\frac{1}{2}}$ or $0{\cdot}25\psi H$, this suggests plasticity indices as low as $\psi = 0{\cdot}1$.

Figure 7 shows the effects of three different surface treatments applied to mild steel. A ground surface (figure 7 (*a*)) gave Gaussian distributions for both the surface

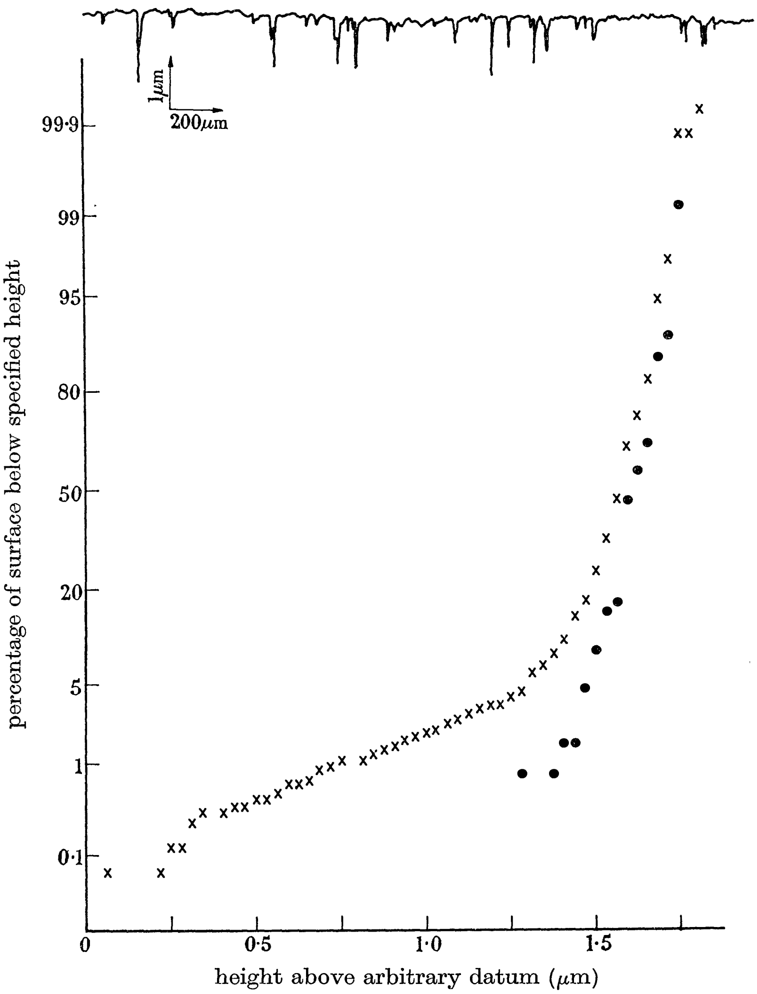


FIGURE 6. Cumulative height distribution of mild steel specimen. Distribution of all heights, ×. Distribution of peaks, ●. This specimen was abraded on 400 grade carborundum paper, then slid against a copper block flooded with oleic acid, at approximately 10 Kg, 130 cm/s for 30 s. Although the distribution is at first sight highly non-Gaussian, in fact nearly 90% of the surface is approximately Gaussian; the surface, with an actual standard deviation of 1·3μm, would behave in contact as if Gaussian with a standard deviation of half this. The profile of the same surface is shown in the upper diagram: the vertical magnification is 200 times the horizontal magnification.

and the peaks. Lightly abrading the ground surface on 600 grade metallographic paper under water preferentially removed the higher parts of the surface, producing a non-Gaussian distribution (figure 7 (*b*)). The mean peak radius changed from 15 to 63 μm, and the plasticity index was reduced from 7 to 2·7. Lightly polishing a ground specimen on a metallographic polishing wheel has a quite different effect (figure 7 (*c*)): the distributions remained almost Gaussian, but, as the profile clearly shows, the small scale roughness was removed from both peaks and valleys. The result was to increase the mean peak radius even more, to 83 μm, and to reduce ψ to 2·5.

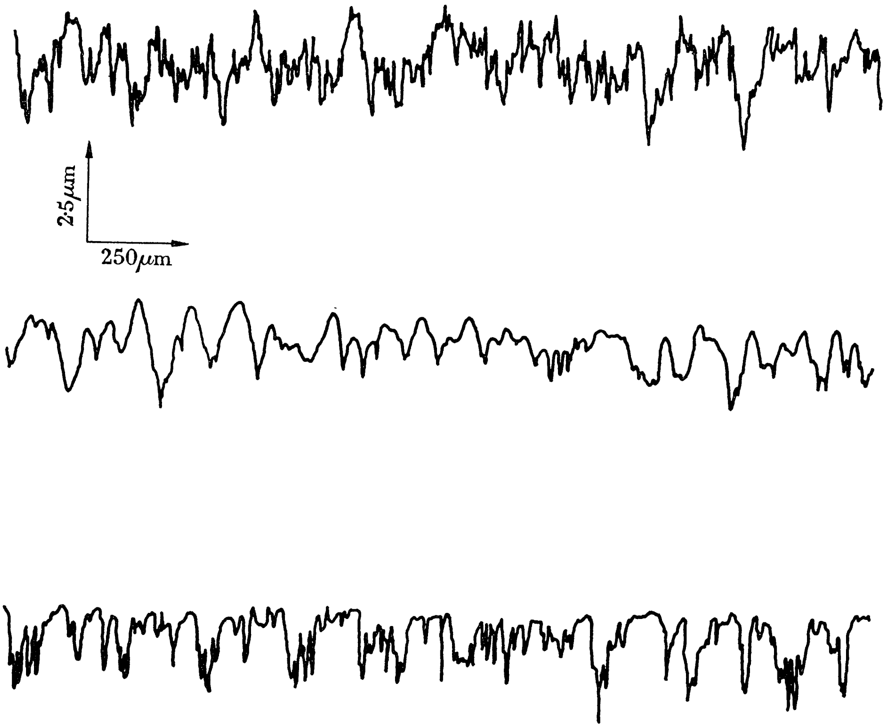


FIGURE 7. Profiles of mild steel specimen after three surface treatments. (*a*) Surface ground only. (*b*) Surface ground and then lightly polished. (*c*) Surface ground and then lightly abraded on 600 paper.

In addition to a distribution of heights of asperities, a surface is also characterized by a distribution of asperity radii, or equivalently, asperity curvatures. The distribution of curvatures is the more convenient, since the data sometimes fit a Γ distribution, in which case the higher moments of the distribution of radii would be infinite. The curvatures are sometimes almost Gaussian, as in figure 8, which shows the results found for a bead-blasted gold specimen.

Certain assumptions are implicit whenever data obtained from a linear profile are used to describe the topography of a surface. Abbott & Firestone referred to cumulative height distribution curves such as figure 5 as 'bearing area curves'. Several authors have recently suggested that these are only 'bearing line curves', and that two such distributions, from perpendicular profiles, must be multiplied together to produce a genuine height distribution of a surface (and thus a reliable value of σ). This is not so. The height distribution would in principle be obtained from an infinite number of closely spaced parallel sections—the usual process of integration over a surface. However, when the profiles are long compared with the surface irregularities they will all contain the same information, so that if, for example, 10 % of one profile lies above a certain height then 10 % of every profile does, and the correct interpretation is that 10% of the surface, not 1%, lies above this height. This is clearly true for randomly structured surfaces: with strongly oriented surfaces it is necessary to add that the profiles must be made so as to include a representative sample of the topography.

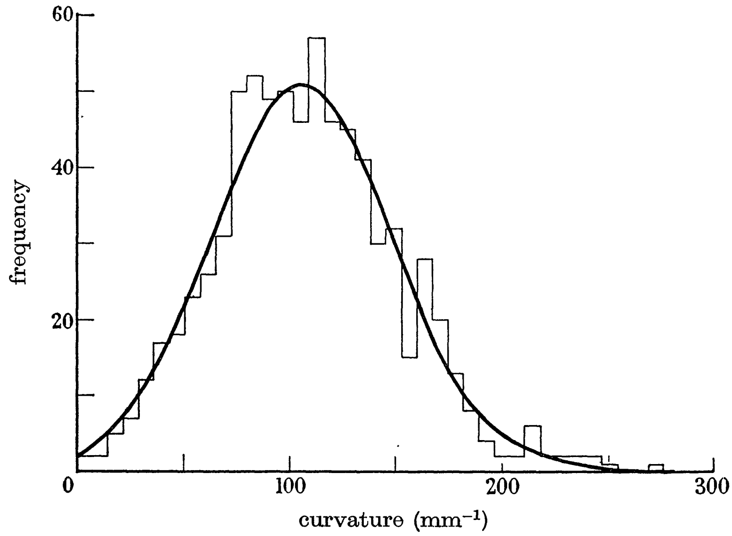


FIGURE 8. Histogram of curvatures at peaks for bead-blasted gold specimen. The continuous curve is the Gaussian with the same mean and standard deviation.

Even though the profiles will in general not cross the summits on the surface the radius of their peaks is a good approximation of that of the asperities. If we consider a single spherical asperity it can easily be shown that the radius of a profile which misses the summit by x is $(\beta^2 - x^2)^{\frac{1}{2}}$. This means that the true radius, β, exceeds the profile radius by a factor of $\sec\alpha$, where α is the maximum slope of the surface at the point represented by the peak of the profile; thus even for a surface with a 25° slope the maximum error will be only 10 %, and the average error will be much less. For a paraboloidal asperity the error vanishes.

A POSSIBLE TOPOGRAPHIC MECHANISM OF WEAR

These results show that plasticity indices ranging from 30 to 0·25 can readily be created by normal techniques; also that abrading and polishing reduce the plasticity index. This suggests that the wearing-in process is the gradual reduction of ψ from an initial 'plastic' value to one in the elastic range. The detailed mechanism of the creation of a wear particle is not relevant here: the hypothesis merely assumes that wear is very much more probable when asperities touch plastically than in purely elastic contacts. Our first attempts to demonstrate this argument were unsuccessful. Although 'elastic' surfaces could readily be made by several fine-polishing processes, it was not possible to generate them by rubbing two 'plastic' surfaces together. Initially the surfaces became polished: the higher parts deformed to leave plateaux separated by the original valleys. But as sliding continued the surface lost its overall flatness and became smooth rolling hills, on a larger scale than the original roughness. This may be attributable to smoothing of adhering wear particles. Carefully selected areas gave plasticity indices as low as 0·5, but these were not representative of the whole surface. The problem was further complicated by surface scoring (roughening), presumably due to abrasion by prows (Cocks 1962, 1964; Antler 1964) or by wear particles.

It occurred to us that sliding a flat specimen against a woven metal grid would still permit the repeated plastic contact between asperities which leads to smoothing, but would minimize damage, by interrupting the process of prow-formation or by removing the wear particles after very short sliding distances.

Three surface-ground mild steel specimens were slid at 100 cm/s under a load of 1·5 Kg against a stainless steel woven grid which was flooded with ordinary machining coolant oil. The grid was changed regularly to avoid the development of wear flats. Figure 9 shows that under these conditions the plasticity index does indeed fall until it reaches the elastic range; and it is clear that the rate of change

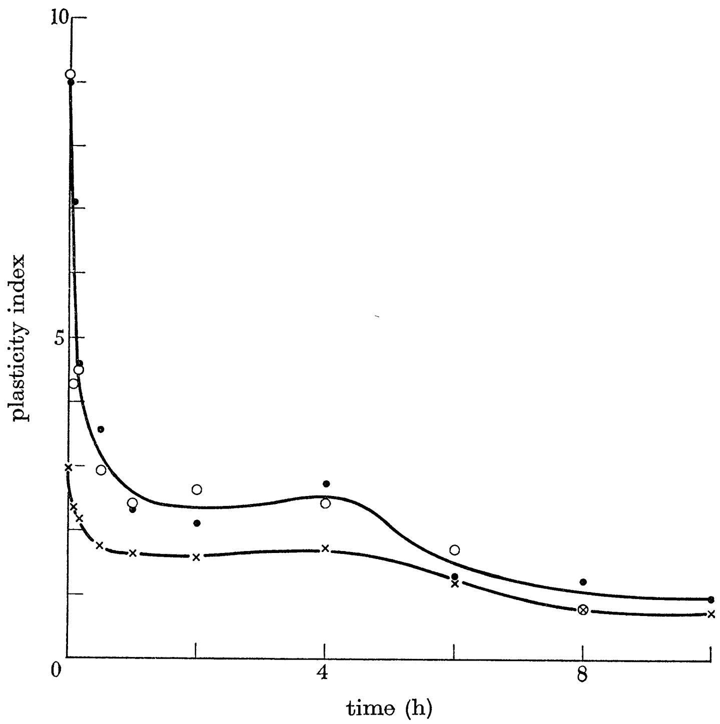


Figure 9. Effect of continued sliding on the plasticity index. The behaviour of three mild steel specimens slid against a stainless steel woven grid at 100 cm/s under loads of 1·5 Kg: ● and ○ show results for two specimens with initial plasticity indices of 9; the crosses show the behaviour of a specimen which had a lower (though still plastic) initial value of ψ.

decreases as the fraction of the contacts which are elastic increases. The manner in which ψ varies during the experiment is much more significant than its absolute value; for the spacing between the height readings determines the smallest peak which can be detected and thus the average radius is to some extent a function of the experimental resolution. During this experiment the standard deviation of the asperity heights, and the c.l.a., of the two rougher specimens fell by a factor of three, and those of the smoother specimen fell slightly: the mean radius of the asperities, on the other hand, increased by much larger factors of 20 and 10 respectively, thus clearly showing the inadequacy of the simple measures of surface roughness.

When the experiment was continued for a further 2 h, the roughening process became dominant and the plasticity index rose again. There was a slight reduction in mean radius, and a much larger increase (threefold) in the c.l.a. and the standard deviation of asperity heights. At the same time wear flats appeared on the grid. It is not clear whether these were responsible for the roughening, or whether a transition from mild to severe wear from some other cause was responsible for both.

Conclusion

The theory developed in this paper leads to a set of relations which give the total real area of contact, the number of microcontacts, the load, and the conductance between two surfaces in terms of the separation of their mean planes. Whereas the separation depends on the nominal pressure (that is, the load divided by the nominal area of contact), the number of microcontacts and the total area of contact depend on the load only. The separation is not very sensitive to the pressure: in fact the mean planes of two similar surfaces in contact are usually separated by 1 to 2 times the standard deviation, or roughly by the centre line average. This means that the average gap between 20 μin. surfaces is, for a wide range of loads, approximately 20 μin. This explains the difficulty of making metal-to-metal gastight seals.

The area of contact and the load depend on the separation in similar ways, so that their ratio is almost constant. This leads to the concept of an 'elastic hardness' by means of which the area of contact can be predicted from the load just as it is for plastic contact using the conventional hardness.

The theory provides a criterion which indicates whether contact will be elastic or plastic: this, the 'plasticity index', is essentially the ratio of the elastic hardness to the real hardness. For most surfaces the deformation mode cannot be affected by changes in the load. It will be elastic if the plasticity index is low, and plastic if it is high. The widespread idea that in general contact is elastic at low loads and becomes plastic as the load increases is wrong. The index, which is equal to $(E'/H)(\sigma/\beta)^{\frac{1}{2}}$, may be regarded as a generalized surface texture parameter, combining both material and topographic properties.

The contact between solids is controlled by two material properties, the plane-stress elastic modulus, and the hardness; and three topographic properties, the surface density of the asperities, the standard deviation of their height distribution, and their mean radius. The commonly quoted parameter 'centre-line-average' is useful only in that it is a loose measure of the spread of asperity heights. In order to

20-2

be able to explore the practical implications of this theory a system has been developed which can measure these new topographic parameters from profilometric observations.

The range of peak radii is very large. Many surfaces have a mean value of 10 to 20 μm; but values of over 500 μm are not unusual. A peak 0·1 μm high may have a radius of 1 mm; the radius need not be of the same order as the height, as sometimes assumed. This of course completely changes estimates of the load a single asperity can carry without yielding. Many common surfaces have a Gaussian distribution of asperity heights. Some types of surface, e.g. bead-blasted ones, are Gaussian to a very good approximation. Others are definitely not Gaussian; however, we have not found any evidence that the non-Gaussian ones fit any other well-known distribution (for example, the extreme value distribution). Even surfaces which have unsymmetrical profiles usually give a straight line on normal probability paper for all but the lowest tenth; and the distribution of their peaks is even closer to Gaussian.

The chief significance of the Gaussian distribution in contact theory appears to be that over a limited range it approximates to an exponential distribution. For the exponential distribution there is exact proportionality between the number of contacts, the area of contact, and the load: thus the average size (and the distribution of sizes) of a contact area is independent of the load. For the Gaussian distribution, among others, these results still hold approximately, and this leads to the idea, which we have discussed elsewhere, that the laws of friction are not the result of material properties, but arise directly from the statistics of surface topography.

Although most common surfaces have plasticity indices well above 1, and thus deform plastically on contact even at the lightest loads, there are several types of surface used in normal engineering with plasticity indices well into the elastic range; a brief survey revealed examples of ψ values ranging from 30 to 0·25. The repeated contact between surfaces, as in normal sliding or in metallographic polishing, tends to reduce ψ. However, the wear debris produced in sliding can reverse this trend, and by damaging the surface can prevent ψ from reaching the elastic range. Subdividing the nominal area of contact reduces the effectiveness of the roughening process, and the gradual decrease of the plasticity index from its initial value in the plastic range to a value in the elastic range is clearly revealed.

We wish to acknowledge the valuable contributions of Mr R. Russakoff who designed and built the topographic analyser, and of Mr R. T. Hunt who carried out most of the experiments. The development of the theory was greatly helped by many discussions with Dr J. H. Tripp.

REFERENCES

Abbott, E. J. & Firestone, F. A. 1933 *Mech. Engng* **55**, 569.
Antler, M. 1964 *Wear* **7**, 181.
Archard, J. F. 1957 *Proc. Roy. Soc.* A **243**, 190
Bickel, E. 1963 *Proc. Int. Conf. Prod. Eng. Res., Pittsburgh.*
Blok, H. 1952 *Proc. Roy. Soc.* A **212**, 480.
Bowden, F. P. & Tabor, D. 1954 Part I, 1964 Part II *Friction and lubrication of solids.* Oxford University Press.
Cocks, M. 1962 *J. Appl. Phys.* **33**, 2152.
Cocks, M. 1964 *J. Appl. Phys.* **35**, 1807.
Greenwood, J. A. 1965 ASME paper 65-Lub-10, to appear in *J. Basic Engng.*
Greenwood, J. A. & Tripp, J. H. *J. Appl. Mech.* (in the press).
Halliday, J. S. 1955 *Proc. Inst. Mech. Engrs* **169**, 777.
Holm, R. 1958 *Electric contacts handbook.* Berlin: Springer.
Miller, J. C. P. 1964 Article 19 in *Handbook of mathematical functions*, National Bureau of Standards, Washington, D.C.
Tabor, D. 1951 *The hardness of metals.* Oxford University Press.
Timoshenko, S. & Goodier, J. N. 1951 *Theory of elasticity.* New York: McGraw-Hill.