ELSEVIER

Tribology International 38 (2005) 786–797

TRIBOLOGY INTERNATIONAL

www.elsevier.com/locate/triboint

The influence of repeated loading, residual stresses and shakedown on the behaviour of tribological contacts

J.A. Williams*

Cambridge University, Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK

Available online 7 April 2005

Abstract

Most tribological pairs carry their service load not just once but for a very large number of repeated cycles. During the early stages of this life, protective residual stresses may be developed in the near surface layers which enable loads which are of sufficient magnitude to cause initial plastic deformation to be accommodated purely elastically in the longer term. This is an example of the phenomenon of 'shakedown' and when its effects are incorporated into the design and operation schedule of machine components this process can lead to significant increases in specific loading duties or improvements in material utilization. Although the underlying principles can be demonstrated by reference to relatively simple stress systems, when a moving Hertzian pressure distribution in considered, which is the form of loading applicable to many contact problems, the situation is more complex. In the absence of exact solutions, bounding theorems, adopted from the theory of plasticity, can be used to generate appropriate load or shakedown limits so that shakedown maps can be drawn which delineate the boundaries between potentially safe and unsafe operating conditions. When the operating point of the contact lies outside the shakedown limit there will be an increment of plastic strain with each application of the load—these can accumulate leading eventually to either component failure or the loss of material by wear.

Keywords: Shakedown; Residual stress; Repeated loading

1. Introduction

The Merriam-Webster on-line dictionary (www.m-w.com) provides as one definition of a machine 'an assemblage of parts that transmit forces, motion, and energy one to another in a predetermined manner'. So with both forces and motion within an assemblage of connected parts, engineers concerned with the design, operation or maintenance of any form of mechanism or machine necessarily have an interest in tribology.

Within such a device, wherever two surfaces are loaded together there will always be some distortion of each of them. These deformations may be very small and purely elastic or they may involve some additional plastic, and so permanent, changes in shape. In the case of non-conformal contacts, whether on the macro- (i.e. component) or micro-scale (i.e. asperity), it is conventional to model the surface stresses as being Hertzian—this is equivalent to supposing that, over the extent of the contact patch, the distribution of pressure between the two surfaces is semi-elliptical. It is worth pointing out that not all contacts either are, or can be sensibly treated as, Hertzian: the idealisations and restrictions implicit in the Hertzian analysis must not be forgotten. Within a Hertz contact, at least in material that is initially free of stress and when traction or friction coefficients are less than about 0.3, the most heavily loaded element of material, and thus the location of first yield, is not actually at the surface but a little way below it, as illustrated in Fig. 1. This region of first plasticity is thus completely surrounded by material which remains elastic. Consequently, the initiation of yield will not be immediately apparent to the superficial observer as the scale of the plastic strain must be of the same order as the elastic limit of the material which, in metals, is only a fraction of a percent. Under repeated application of the load, as for example occurs in a rolling contact, the fatigue life of the component (or perhaps the wear rate of the surface) may be expected to depend upon the progress of such plastic deformation. It is possible for a load which generates plastic flow on its first application to induce on its removal a system of residual protective

* Tel.: +44 1223 332625; fax: +44 1223 332662.
E-mail address: jaw@eng.cam.ac.uk.

0301-679X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.triboint.2005.02.006

Nomenclature

a	semi-contact width of Hertzian line contact
b, d	beam dimensions
k	shear yield stress of the material
E	elastic modulus
E^*	contact modulus ($=E'/2$)
M	bending moment
M_{el}	elastic bending moment
M_{pl}	plastic bending moment
N	asperity density
P	normal load
p	normal pressure at the contact
$\bar{P}$	normalised load intensity
P_s	load intensity at shakedown
p_0	maximum contact pressure
p_s	contact pressure at shakedown
Q	traction or friction load
q	shear traction within the contact
R	asperity or component radius
Δu_{xx}	surface displacement
$\bar{U}$	entraining velocity
U, W	non-dimensional parameters in lubricated sliding
w	load per unit length of contact
ν	Poisson's ratio
σ_y	uniaxial yield stress
μ	coefficient of sliding friction
σ	rms roughness
σ_{ij}	stresses
ρ_{ij}	residual stresses
α_{ij}	back stresses
Ψ_s	plasticity index
η_0	viscosity
τ	shear stress

stresses which will allow the same load, when re-applied, to be carried entirely elastically.

This process is referred to *shakedown* and the maximum load for which it occurs is known as the *elastic shakedown limit*. Under more severe load conditions no shakedown state may be possible so that plastic deformation will continue to take place with every subsequent imposition of the load. When shakedown does occur it may do so as rapidly as the second application of the load, or it may take many load cycles to become fully established—such variations depend on the particular material properties of which the surface is made and may be influenced by local environmental conditions, thermal gradients, etc.[1]

The application of the principles of shakedown to the repeated Hertzian loading of a half-space which possesses the sort of mechanical properties typical of real, as opposed to idealised, engineering materials is not straightforward. The fact that the stress field beneath such a loading, particularly if there is an element of surface traction, is itself complex can complicate the essential features of the argument. However, the principles can be illustrated by reference to the simpler, and more familiar case of a beam loaded so as to be subject to pure bending.

2. Shakedown of a simple structure

Consider a beam of rectangular cross-section say breadth b by depth d, Fig. 2(a), made of a material which is *elastic perfectly-plastic* with yield stress σ_y, Fig. 3(a). The beam is initially stress free. In the figures, in order to save space, only the top part of the beam, i.e. $0<y<d/2$ is considered; the lower half is just the anti-symmetric image. When put into pure bending we make the usual geometric idealisation that strains are sufficiently small for plane sections to remain plane, so that the magnitude of the longitudinal strain varies linearly with distance from the unstrained or neutral axis. If the applied bending moment is insufficient to cause yield at outer fibres of the beam then, since the material is linear elastic, the axial stresses in the material likewise vary

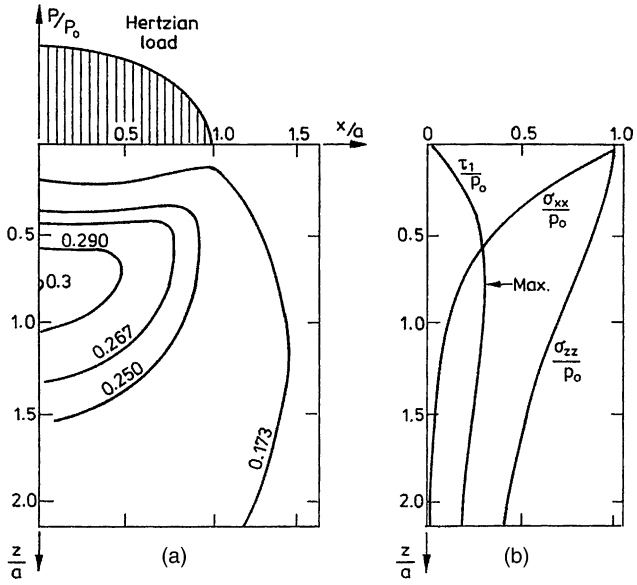

Fig. 1. Stresses beneath a Hertzian line load. (a) Contours of maximum values of the shear stress on an x–z plane. The most heavily loaded element of material occurs at a depth of $0.78a$. (b) Distribution of direct stresses σ_{xx}, σ_{zz} and τ_1 the maximum shear stress equal to $|\sigma_{zz}-\sigma_{xx}|/2$ all normalised by the peak hertz stress p_0 along the axis of symmetry Oz.

[1] In what follows the material is modelled as a continuum described by the simplest constitutive equation which captures its essenbtial mechanical behaviour. For an introduction into the structural implications the reader is referred to Rie K-T, Portella PD. Low cycle fatigue and the elasto-plastic behaviour of materials. Amsterdam: Elsevier; 1998.

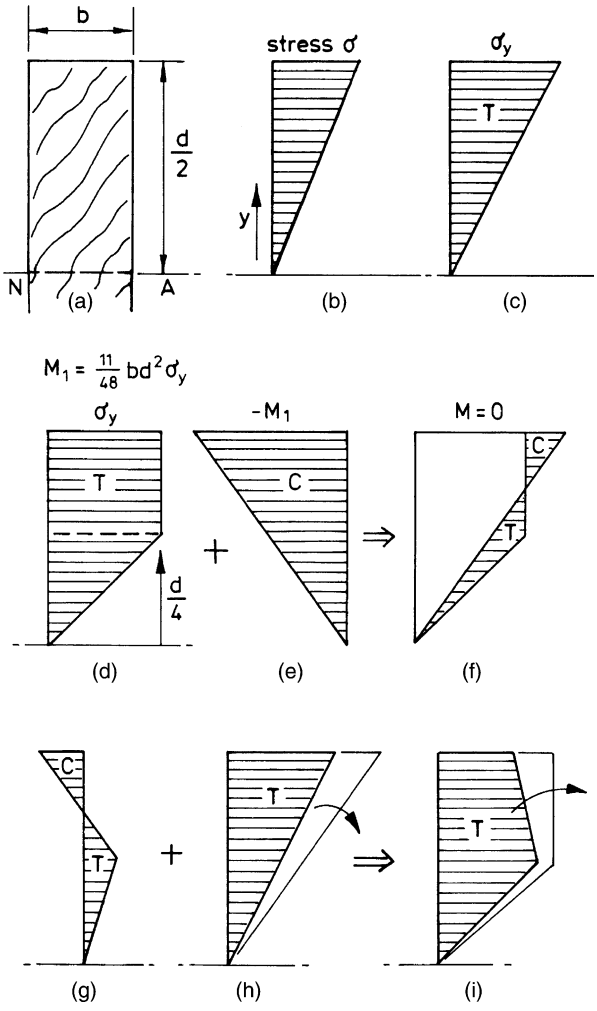

Fig. 2. Elastic–plastic loading of a beam of rectangular cross-section $b \times d$ in bending; only conditions above the neutral axis are shown, those below are the asymmetric reflection. (a) Dimensions. (b) Stress distribution under simple bending, fully elastic regions in tension T, those in compression C. (c) At the limit of elastic behaviour M_{el} the outer fibre stress just reaches σ_y (d) when the plastic zone extends to 50% of the beam area the applied moment is M_1. (e) Unloading can be thought of as the application of a reverse moment of magnitude M_1. (f) and (g) Residual stresses, sum of (d) and (e). (h) Reapplication of M_{el}. (i) Fully elastic conditions are now possible for $M_{el} < M < M_1$.

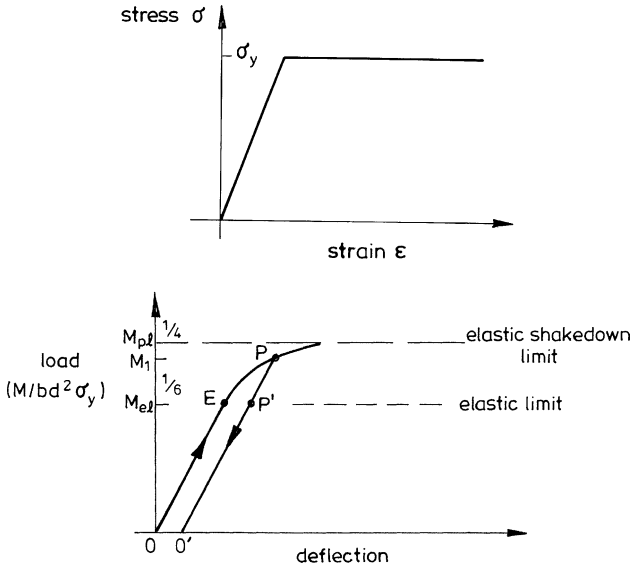

Fig. 3. (a) Stress–strain characteristic of an elastic-perfectly plastic material with a tensile yield stress σ_y. (b) The load deflection behaviour of the beam of Fig. 2. M_{el} is the bending moment at which the outermost fibres of the beam reach yield and M_{pl} that at which the whole section is plastic. M_1 is an intermediate value, here chosen so that the plastic zone extends over the region $|y| \geq d/4$ so that $M_1/bd^2\sigma_y = 11/48$.

linearly with coordinate y as illustrated in Fig. 2(b). On a plot of load versus deflection (perhaps the relative rotation of the ends of the beam) the behaviour of the beam is perfectly linear as shown by the line OE in Fig. 3(b). If the load is removed the beam returns to its original stress free condition. At some value of bending moment, say M_{el}, the stress in the outer fibres of the beam just reach the value σ_y shown as Fig. 2(c); the beam has reached its elastic limit which is say at point E on Fig. 3(b). The value of M_{el} is $M_{el} = bd^2\sigma_y/6$.

If we now increase the applied bending moment, a zone of plasticity, in which the longitudinal stress is equal to σ_y, grows from the outer surfaces of the beam towards the neutral axis. When the whole of the beam reaches this condition a plastic hinge is formed and the beam collapses, i.e. continues to deform with no associated increase in load. This occurs at a load of $M_{pl} = bd^2\sigma_y/4$. But, suppose we only take the beam part of the way from the limit of elasticity towards full plasticity, say to where the plastic zone extends to $y = d/4$. Then the necessary bending moment is $M_1 = 11bd^2\sigma_y/48$, i.e. between M_{el} and M_{pl}, and the distribution of stresses is as shown in Fig. 2(d). The state of the beam is now at point P on Fig. 3(b); the fact that some of the beam is behaving plastically leads to an increase in the deflection of the beam to values beyond those that would arise simply from linearity.

Now we unload the beam. On Fig. 3(b) the beam unloads along an elastic line parallel with the initial elastic line from P to O', the intercept on the deflection axis OO' representing the resultant remaining deflection at zero load. This must still be 'small' even though plasticity has been involved, because the deformation of the central part of the beam is limited to small elastic strains. We can infer what is happening inside the beam by adding to the stress distribution of Fig. 2(d) the stresses involved in unloading, equivalent to applying a negative M_1, i.e. the stresses of Fig. 2(e). We can suppose this to be elastic and just check that there is no reverse plasticity. Adding Fig. 2(d)–(e) is the same as finding the difference between 2(d) and the negative of Fig. 2(e) as in Fig. 2(f) which we can plot out as Fig. 2(g) just bringing the zero stress line vertical. The outer fibres of the beam on its top surface are left with a residual compressive stress while there is a compensating region of tensile stress nearer the centre; the whole beam is of course in overall equilibrium.

Now apply the positive bending moment again, gradually increasing from origin O. When the applied moment is equal to M_{el} the internal stresses are the sum of Fig. 2(g) and (h) to give Fig. 2(i). There is no zone of plasticity. On first loading with M_{el} we got to Fig. 2(c), on the second application to Fig. 2(i). The outer fibres are well away from plasticity because they started now with a residual compressive stress which the applied load has to overcome. On Fig. 3 we have moved from O' to P'.

Now increase the applied moment slowly to M_1. The effect on the internal stress distribution is to gradually increase the load on the outer region of the beam—but *there is no plasticity* until the applied bending moment reaches the value M_1 and we move from P' to P on Fig. 3(b). We can cycle quite happily elastically from O' to P on Fig. 3(b). Only if the applied bending moment exceeds M_1 do we start to get plastic flow again. Because of the introduction of the residual stress field the limiting elastic load has increased from the associated with operating point E to that of P. The beam has exhibited shakedown. The stress–strain characteristic of the material has not changed but the load deflection characteristic of the component has. It is very important to draw the distinction between the behaviour of the *material* Fig. 3(a) and that of the *component* Fig. 3(b). These will only be congruent if every element of the component sees exactly the same load history. This would be true for a simple tension specimen but is not the case for a bending beam or, for that matter, a loaded surface. In principle, this argument is acceptable for any value of bending moment between M_{el} and one infinitesimally below M_{pl}. In this case the plastic collapse load M_{pl} is an upper bound on the shakedown load—it represents the shakedown limit and the beam has shaken down in one application of the load.

To couch this procedure in terms of a fatigue test it is probably easiest to think of the beam as being subject to two applied bending moments: the first steady and of value $1/2M_1$ and the second a cyclic load of magnitude $\pm 1/2M_1$. Thus, since the beam has shaken down by the time its history has reached the point P in the load path it can be thought of as having done so in one quarter of a fatigue cycle. This is a characteristic feature of the shakedown of structures made of an elastic-perfectly plastic material, i.e. a material which, beyond the elastic limit, deforms under a constant flow stress (it is also true of a material which exhibits isotropic hardening). The effect is solely due to the generation of protective residual stresses.

3. The shakedown limit and the shakedown theorems

In general, one of two approaches can be used to obtain elastic shakedown limits. The first is numerical. The component is modelled as a finite element mesh and, using a suitable elastic–plastic code, the load is applied so that the internal stresses and components of deformation of each element of the mesh can be determined. The load is then removed (usually by applying an appropriate negative loading at the boundary) and thus the residual deformation and stresses within the bulk evaluated. In the next application of the load the effect of these are thus taken into consideration in evaluating new up-dated values. The process is then repeated and after (perhaps) many cycles a steady state reached in which deformation and residual stress no longer change from one cycle to the next—this implies that elastic conditions have been achieved.

This process has been applied to tribological problems. One component is modelled as a curved point or line contact and the other as a half-space across which the first is traversed. By repeating the procedure for increasing levels of load, shakedown limits for the contact can be established [1,2]. However, since the analysis involves elastic–plastic finite element calculations and approaches the steady state in an iterative manner, considerable computational effort is required. The solutions also suffer from the draw back of being non-analytic so that it is not easy to predict the effect on the shakedown load of changes in the material or geometric parameters. More recently computational methods of greater efficiency which approach the steady-state more directly have been developed although they are still essentially numerical and so non-analytic in nature [3–7].

A second quite different approach, and which we follow here, makes use of the shakedown theorems of the theory of plasticity to obtain upper and lower bounds to the shakedown limit. The upper bound establishes a maximum possible value for the shakedown load and the lower bound a corresponding minimum value. Even though the exact shakedown load may not be known it must lie between these two limits. By refining these estimates it is possible to bring the maximum and minimum values sufficiently close to obtain a useful approximation to the actual shakedown limit. This technique has the advantage of being predominantly analytic and thus indicating the importance of the variables within the problem; since it approaches the steady state directly and uses only elastic stresses, it requires much less computational effort.

For a rigid-perfectly plastic solid the upper and lower bounds to the collapse load, i.e. the load at which material undergoes plastic flow in the *first* application of load, is provided, respectively, by the familiar structural statical and kinematic theorems (for example, see [8,9]). The corresponding and appropriate theorems for an elastic-perfectly plastic solid subjected to *repeated* loading have been obtained by Melan [10] and Koiter [11]. They can be stated as follows:

(i) Melan's Statical Shakedown Theorem: 'If any system of self-equilibrating residual stresses can be found which, in combination with the stresses due to the repeated load, do not exceed yield at any time, then elastic shakedown will take place'. The maximum load which, together with the 'true' distribution of residual

stress, just touches yield gives the exact shakedown limit. Any other distribution provides a lower bound to the shakedown limit. In practice many possible residual stress distributions are examined and the one providing the highest lower bound is chosen.

(ii) Koiter's Kinematical Shakedown Theorem: 'If any kinematically acceptable mechanism of incremental plastic collapse can be found in which the rate of work done by the elastic stresses due to the load exceeds the rate of plastic dissipation, then incremental collapse will take place. The ratio of the work done by the elastic stresses to the plastic dissipation has a maximum in the 'true' mechanism of collapse, so that any other mechanism gives an upper bound to the shakedown limit. We might examine several different possible 'collapse mechanisms' and chose that with the lowest associated collapse load.

Application of these theorems thus provides bounds to the true shakedown load: the structure can be no weaker than the highest lower bound or stronger than the lowest upper bound. Of course if these two bounds merge, then the exact limit has in fact been established.

4. Shakedown in sliding or rolling line contacts—shakedown maps

Tribological contacts are conventionally characterised as either *line* contacts (typified by a cylinder rolling or sliding across on a flat) or *point* contacts (corresponding to a sphere on flat). While both may undergo the phenomenon of shakedown through the effects of material hardening and the generation of residual stresses, it is clear that in the three-dimensional case an additional mechanism, that of increased conformity between the two bodies as the sphere generates a groove in the counterface, can come into play. In what follows we do not consider this geometric change—the interested reader is referred to Kapoor and Johnson [12]—but restrict ourselves to essentially two-dimensional line contact and consider first the effect of friction at the interface.

4.1. Frictionless contact

The contact stresses due to the pressure of a rigid cylinder on a semi-infinite solid follow from the Hertz theory. The normal pressure $p(x)$ at the interface is distributed semi-elliptically, i.e.

$$p(x) = p_0\sqrt{1 - x^2/a^2}. \quad (1)$$

The semi-contact width a and the maximum contact pressure p_0 are given by

$$a = \sqrt{4PR/\pi E*} \text{ and } p_0 = \sqrt{PE*/\pi R}, \quad (2)$$

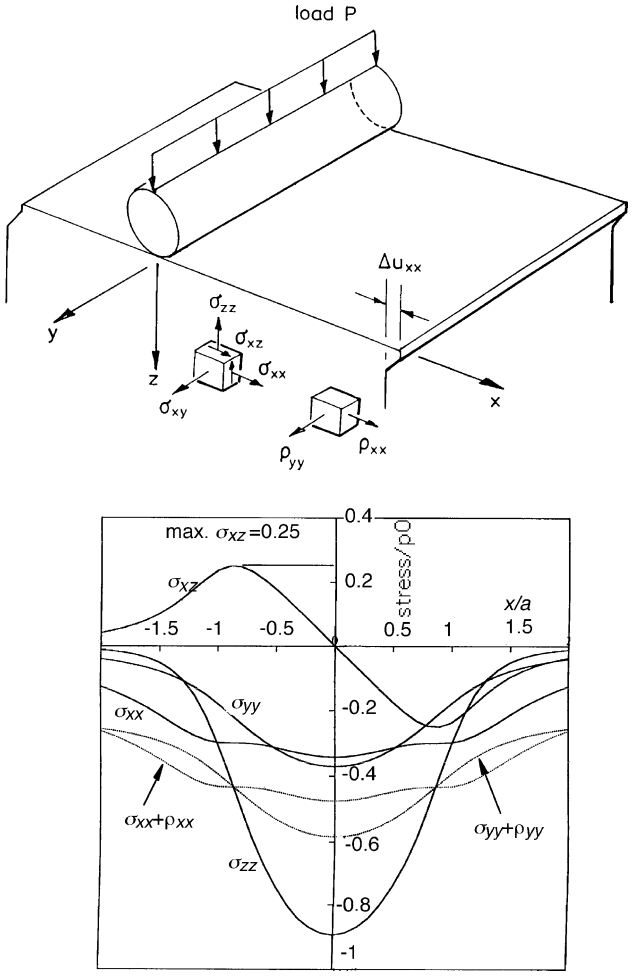

Fig. 4. (a) A travelling Hertzian line load. σ_{xx}, σ_{yy}, σ_{xz} is stress system due to the load P; ρ_{xx}, ρ_{yy} are residual stresses. Δu_{xx} is the surface translation applicable in the application of Koiter's kinematical theorem. (b) Solid curves: the internal stresses beneath such a distribution at a depth of $z=0.5a$ which comprise a shear stress σ_{xz} and direct compressive stresses σ_{xx} and σ_{yy}. Dotted curves: stresses after shakedown, i.e. with the addition of the residual stress components ρ_{xx} and ρ_{yy}.

in which R is the radius of the cylinder, P is the load per unit length, as indicated in Fig. 4(a) and the contact modulus $E^* = \{(1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2\}^{-1}$. The internal stresses beneath such a distribution can be readily computed, for example, Poritsky [13], Smith and Liu [14], and Sackfield and Hills [15]; they comprise a shear stress σ_{xz} and direct compressive stresses σ_{xx} and σ_{yy}. These are shown as the solid curves in Fig. 4(b) for conditions at a depth of $z=0.5a$; the reason for this choice of depth will be clear when we examine the application of Melan's theorem. In a frictionless sliding contact or a free rolling contact, i.e. one in which no traction stresses are applied to the roller at the interface, these curves can be thought of as expressing the loading history of a material element as the load completes one pass [16].

The value of the load intensity, i.e. the numerical value of p_0/k where k is the yield stress of the material in simple shear, at which some element of the material is first loaded beyond the elastic limit under a Hertzian line contact

depends to some extent on the yield criterion adopted. For example, using the Tresca maximum shear stress criterion (see, for example [8, p. 38]) it is clear from the stress profiles in Fig. 1 that the maximum shear stress is of magnitude $0.3p_0$ and occurs at a depth of $0.78a$. Thus Tresca would imply that for first yield $0.3p_0 = k$, i.e. $p_0 = 3.3k$. The von Mises (or 'J_2') criterion [8, p. 47] involves the third out-of-plane principal stress and thus will be influenced by the value of Poisson's ratio ν: taking this as 0.3 leads to first yield at depth $0.7a$ below the surface and a critical value of p_0 equal to $3.1k$.

Application of Melan's theorem requires, in addition to the set of stresses illustrated by the solid lines in Fig. 4(b), a system of self-equilibrating residual stresses which we shall call ρ_{xx}, ρ_{yy}, etc. Clearly, ρ_{zz} must be zero (since the upper surface is stress free) as (by symmetry) so must ρ_{xy}, ρ_{yz} and ρ_{xz}. The residual stress component ρ_{yy} which varies with depth z can be made the intermediate principal stress so that, if yield is not to be exceeded by the applying the set Tresca criterion

$$\frac{1}{4}\{(\sigma_{xx} + \rho_{xx}) - \sigma_{zz}\}^2 + \sigma_{xz}^2 \leq k^2. \tag{3}$$

This condition clearly cannot be satisfied if σ_{xz} exceeds k but can *just* be satisfied with σ_{xz} equal to k if we choose to make $\rho_{xx} = \sigma_{zz} - \sigma_{xx}$. The limiting conditions for which shakedown is just possible occurs at the point in the solid where the value of σ_{xz} is a maximum. In frictionless contact $(\sigma_{xz})_{\max} = 0.25p_0$ at $x = \pm 0.87a$ and $z = 0.5a$, which therefore gives a *lower bound* to the shakedown limit p_s such that

$$p_s/k \geq 4.00. \tag{4}$$

The residual stresses at this point are then $\rho_{xx}/p_0 = -0.134$ and, assuming a value of Poisson's ratio of 0.3, $\rho_{yy}/p_0 = -0.213$; addition of these to the travelling set of stresses gives those plotted as dotted lines in Fig. 4(b).

To apply Koiter's kinematical theorem to provide an upper bound on the shakedown limit we must postulate a kinematically acceptable mechanism of incremental collapse. Suppose that this is simple plastic shear along the plane parallel to the outer surface of the solid. If the increment of plastic deformation is Δu_{xx} then the work done by the elastic stresses is $\sigma_{xz} \times \Delta u_{xx}$ and the internal work dissipation is $k \times \Delta u_{xx}$. An upper bound on the shakedown load is thus found by equating these two quantities, i.e. $\sigma_{xz} = k$; but since the maximum value of σ_{xz} is $0.25p_0$ it follows that $p_0 = 4k$ is the optimum upper bound on the collapse pressure, i.e. that

$$p_s/k \leq 4.0. \tag{5}$$

Since, in this case, the lower and upper bounds are identical they describe the true collapse load.

The practical importance of shakedown can now be readily appreciated by comparing the pressures, and hence the loads, for first yield to those required in the steady-state when the residual stress field has become established. Using the von Mises criterion, p_0/k for first yield is 3.1 while the shakedown limit or critical figure is 4.0. However, since the value of the line load P is proportional to $(p_0)^2$, the ratio of the two corresponding values of applied load, say P_Y and P_S, are such that

$$\left\{\frac{P_S}{P_Y}\right\} = \left\{\frac{4}{3.1}\right\}^2 = 1.66. \tag{6}$$

A load 66% greater than that which will cause yield on its first application can be carried safely, that is without subsequent yield, in situations of repeated loading. The appreciation of this very substantial increase in load capacity brought about by shakedown has had important practical and economic implications in the specification of loading duties in industries such as rail transport and rolling element bearings.

4.2. Influence of friction

In a sliding contact in which friction acts (and so in which by definition friction must be limiting) the frictional traction $q(x)$ within the contact area will be given by

$$q(x) = \mu p_0 \sqrt{1 - x^2/a^2}. \tag{7}$$

The subsurface stress field caused by this traction is likewise known [13–15] and can be used to establish the value of p_0/k for first yield using either Tresca or von Mises in much the same way as for frictionless sliding. As the traction stress rises so there is a reduction in the critical value of p_0. For low values of the coefficient of friction (specifically 0.25 for Tresca and 0.3 for von Mises) the yield point is first reached at a point in the material beneath the contact surface. For larger values of μ yield first occurs *at* the contact surface. The fall in p_0 is relatively modest when $\mu < 0.3$ but becomes more marked thereafter. This variation can be conveniently displayed on a 'shakedown map' which plots critical values of p_0/k versus the friction or traction coefficient μ: such a plot is shown in Fig. 5 in which the elastic limit, or first yield line is shown dotted, curve A.

In the case of *repeated* loading, shakedown is again possible just as in frictionless sliding and can be investigated by superposing the subsurface tractive stress field due to Q, i.e. $\int q(x)\mathrm{d}x$, on the stresses due to the normal load P and considering an appropriate residual stress field ρ_{xx}, ρ_{yy}, etc. The criteria are just as before; shakedown is governed by the maximum value of σ_{xz} in the field. For values of μ (i.e. Q/P) less than a little more than 0.3 the maximum value of σ_{xz} occurs at a point beneath the surface, somewhat shallower than in frictionless contact. When μ exceeds 0.3 then $(\sigma_{xz})_{\max}$ occurs at the surface itself with a value μp_0. Hence the shakedown limit, by Tresca, is

$$p_s/k = 1/\mu. \tag{8}$$

The critical values of p_0/k from such an analysis are also displayed on the map of Fig. 5 as the chained line (B);

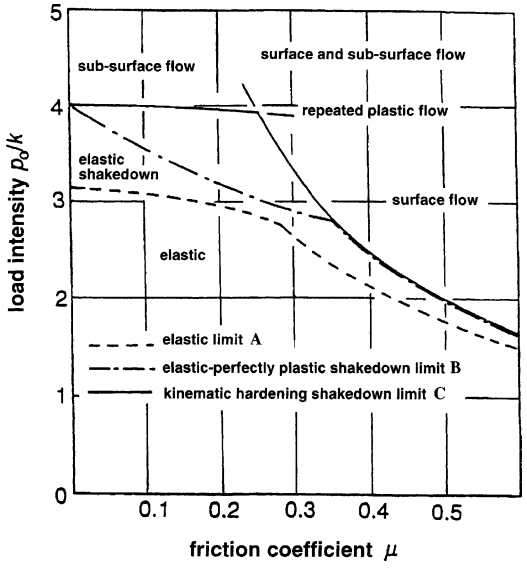

Fig. 5. Shakedown map for line contact in which the maximum allowable Hertz pressure (normalised by yield stress in shear) is plotted against local coefficient of friction. The region below each of the curves can be considered 'safe'. Curve A illustrates the most conservative design in which the elastic limit is never exceeded: curve B shows the effect of allowing for shakedown in an elastic-perfectly plastic material and curve C the additional performance available if the material undergoes kinematic hardening.

the interval between this and the dotted curve thus represents the practical benefit of the shakedown phenomenon in an elastic-perfectly plastic material.

Friction in a sliding contact must, by definition, be limiting. If Q does not reach its limiting value, as might occur in a rolling contact between two elastic solids, then the zone of contact contains both 'sticking' and 'slipping' regions, and, furthermore, the symmetry of the solution is disrupted [17]. Nevertheless it is still possible to generate shakedown maps and for details of these the reader is referred to the papers by Johnson [18,19].

4.3. Effects of strain hardening

Most materials strain harden. Plastic yielding during the early cycles of loading raises the effective yield stress in subsequent load applications, thereby promoting elastic shakedown. To be quantitative, a strain hardening model must be chosen and married with the techniques described above. The simplest model is that of isotropic hardening in which it is supposed that the yield surface when plotted in stress space can expand but not change its location or its shape; all that happens is that with each load application the stress axes are effectively re-scaled. Unfortunately, this is a poor model of real material behaviour under cyclic loading (it necessarily leads to shakedown in a quarter of a load cycle) although an element of isotropic hardening may be usefully retained in more sophisticated hardening models.

The next simplest model is that of kinematic hardening in which the yield surface is allowed to move in stress space but without any change in shape or size: the displacement of the centre of the locus is sometimes designated as a 'back-stress' which has components α_{xx}, α_{yy}, etc. This model is a much better match to the properties of many metals, it can, for example, model the Bauschinger effect (i.e. a reduction in the value of the compressive yield stress after plastic strain in tension—see [8, p. 28]). A further division can be made between so-called linear and nonlinear kinematically hardening materials. In a linear model the hardening rate is uniform whatever the level of mean stress within the loading cycle, whereas in a nonlinear hardening law the response of the material is influenced by this parameter. This distinction is not necessary for the purposes of evaluating elastic shakedown limits; however, it is important in the responses of the material above shakedown.

In the case of a kinematically hardening material an additional shakedown theorem due to Ponter [20] can be invoked; this allows for the displacement of the centre of the yield locus. Shakedown now involves both residual stresses ρ_{ij} (this subscript notation implies the set of stresses ρ_{xx}, ρ_{yy}, etc.) and the back-stresses α_{ij}. The sum of these can be thought of as an effective residual stress ρ^*_{ij} which can be used in Melan's statical theorem to find a lower bound for the shakedown limit. For any material element ρ^*_{ij} must remain constant through the loading cycle, however, unlike the 'true' residual stress it need not satisfy the equations of equilibrium and this simplifies its application although it is does not permit its division into ρ^*_{ij} and α_{ij}. In the problem of repeated Hertzian line contact this means that the stress components ρ^*_{zz} and ρ^*_{xz} need not be zero (whereas clearly both ρ_{zz} and ρ_{xz} are zero) and are permissible in addition to ρ^*_{xx} and ρ^*_{yy}. Shakedown limits found in this way [19] have been shown as the solid line (curve C) in Fig. 5. If the operating point of a sliding contact plotted in this figure lies below curve A (the elastic limit) then no element of material reaches the yield point. In the space between curves A and B a perfectly plastic material will yield initially but will achieve elastic shakedown in the steady state. The elevation of curve B above curve A indicates the contribution of residual stresses to shakedown. Load conditions between B and C will only lead to shakedown if the material is capable of kinematic hardening. The elevation of curve C above B indicates the contribution of this form of hardening to shakedown. If the contact is loaded so that its operating point is above curve C in this plot then plastic deformation will occur in each cycle of loading: linear kinematic hardening leads to closed cycles of plastic deformation and nonlinear hardening to plastic ratchetting.

5. Beyond shakedown

The régimes of response for a component (of which a surface might be an example) are illustrated in Fig. 6. If the loads are sufficiently small for no element of material to reach yield, then the response of the structure will also be

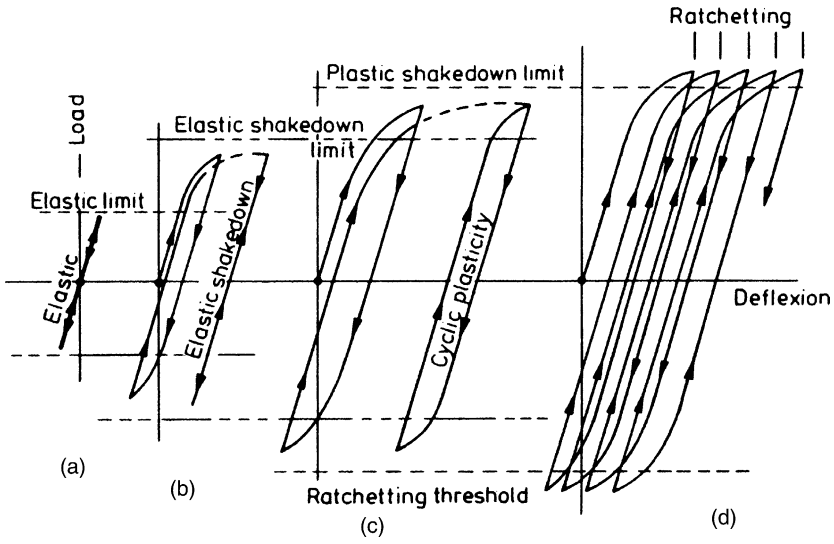

Fig. 6. The different forms of structural response to cyclic loading: (a) perfectly elastic, (b) elastic shakedown, (c) plastic shakedown or cyclic plasticity, (d) incremental collapse or ratchetting.

entirely elastic throughout the load cycle, régime (a). Above the structural elastic limit, plastic flow will be encountered somewhere within the material in at least the first application of the load. However, because of the development of residual stresses, the steady cyclic state may still entirely elastic and this is the region we have referred to as elastic shakedown. At higher loads, each cycle of loading leads to elements of both elastic and plastic deformation. If the steady-state strain cycle is *closed*, so that the regime is one of cyclic plasticity, then it can be said that *plastic shakedown* has been achieved, this is illustrated in régime (c). However, under some circumstances, alluded to above, each cycle of load may generate both reversing and uniaxial components of strain; this is the process known as either incremental collapse or ratchetting failure and is illustrated by régime (d) in Fig. 6: the implications of this régime for the mechanisms of wear and surface degradation have been considered [20].

A contact whose operating point lies above the shakedown limit in an appropriate map, within the ratchetting or incremental collapse regime, is liable to premature failure either by fracture or by wear because of the element of plastic deformation that accompanies each load cycle. Put the other way round, we should expect very long lives from contacts whose operating point lies within the performance envelope represented by the shakedown limit. The stresses in the heavily loaded non-conformal contacts characteristic of rolling element bearings, continuously variable power transmission systems and between cams and followers or individual gear teeth are amongst the highest encountered in mechanical engineering and so it is not surprising that it is in these areas that this sort of analysis has, and continues to have, its greatest impact. A truly conservative design would keep all these stresses below the initial yield limit, i.e. below curve A in Fig. 5, but commercial and market forces will inevitably demand that specific loadings are increased and that the very real increase in component efficiency represented by the interval between curves A and C in Fig. 5 is exploited.

The surface engineering of rolling and sliding machine components is now standard industrial practice. Techniques of diffusion or impregnation improve the wear resistance principally by increasing the hardness—and thus the shear yield stress—of the material at and near the bearing surfaces while leaving the core material relatively soft and tough. Advances in physical and chemical coating technology enable hard ceramic layers to be deposited in particularly vulnerable areas. Fig. 7 [21–23] is a plot of the shakedown limiting pressure p_s (normalised by the core yield stress k_2) in a half-space whose free surface has been hardened to shear stress k_1 in a layer of depth h. Because dimension h is likely to be only of the order of microns, shakedown is now operating at the scale of asperities so that dimension a is

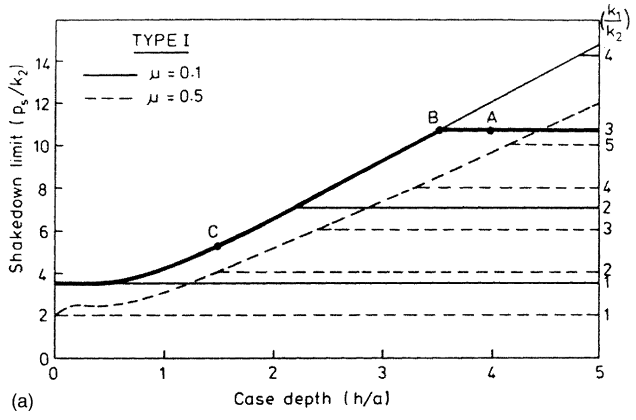

Fig. 7. Shakedown limits for a half-space which has been surface hardened or coated to a depth h with material of shear yield strength k_1. The shakedown pressure is p_s and k_2 is the yield stress of the core. Dimension a is a measure of the size of a representative asperity contact.

the half width of a typical asperity contact and is thus related to surface roughness. Point B indicates that for the case $k_1/k_2=3$ the full increase in p_s can only be achieved if the depth of hardening is at least three and a half times the asperity contact dimension. If the case depth is only of the order of dimension a then hardening the surface has a negligible effect on the shakedown pressure which, for a contact with a coefficient of friction of 0.1, is equal to $3.65k_2$. It is possible that during service mild wear will gradually reduce the thickness of the hardened layer, so moving the operating point of the contact along the curve BC; if the shakedown pressure falls below the actual service pressure then a transition to a much more severe wear rate might be expected. The dashed curves show the effect of increasing the surface friction or traction coefficient to the value 0.5.

6. Some applications

6.1. Rolling element bearings

The first publication of a shakedown in quite this form was in 1963 [24] at a symposium on 'Fatigue in Rolling Contact' which was a milestone in the field for this and a number of other reasons. Several papers at the Symposium provided lubricated rolling contact fatigue data and fatigue limits (i.e. contact stress levels at which failure did not occur during the test) were subsequently extracted and displayed as plots of p_s/k_0, i.e. the maximum Hertz stress as a multiple of shear yield stress, versus the non-dimensional parameter U/W^2 which is equivalent to the group $\bar{U}\eta_0 E'R/w^2$. This is a measure of the distribution of the contact load: a small value corresponding to low speed and high load (for which the pressure distribution will be essentially Hertzian) while high values of this parameter, i.e. high speed and light loads, approach the case of undeformed rollers. A film of lubricant alters the Hertzian distribution of pressure and thus in principle changes the shakedown load, however, this effect is negligible up to values of $U/W^2 \approx 10^{-2}$. Fig. 8, adapted from [25], is the plot in question from which is can be seen that the endurance limit observed experimentally was about half that which is in principle achievable utilising the full shakedown potential. For a line contact this would suggest that an increase in load capacity of 4 was possible while for a point contact where $p_0 \propto \text{load}^{1/3}$ the factor would be 8.

At the time of the experimental work reported in 1963 the explanation for the for the relatively low stresses at which rolling contacts were observed to fatigue involved the effects of subsurface non-metallic inclusions in the steel. These acted as local stress-raisers from which cracks could be generated which subsequently propagated to the surface generating a characteristic arrow-shaped spall [26]. In the years that followed the bearing industry has made strenuous efforts to improve the cleanliness of the steel, principally AISI 52100, that is used to manufacture both the bearing races and rolling elements. Vacuum remelting became much more common and the defect density has correspondingly fallen.

The impact that this has had on the performance is graphically illustrated in Fig. 9(a) which shows the evolution over the last 50 years or so of a double row spherical roller bearing from a major manufacture. All these bearings have or had a similar specification as far as load and speed is concerned of the order of 490 kN at 2000 rpm. In that period the mass of the bearing has reduced by a factor

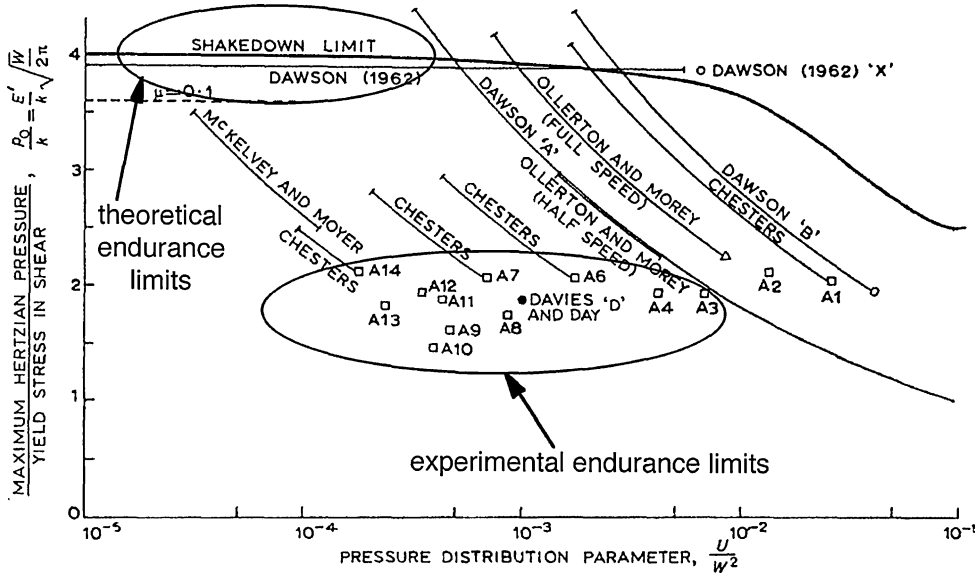

Fig. 8. Data from Fatigue in Rolling Contact a symposium organised by the Institution of Mechanical Engineers in London in 1963 [26]. The intensity of the bulk contact stress for several of the tests (expressed as the ratio of the maximum Hertz pressure to the yield stress in shear) is plotted against the parameter $U/W^2=\eta_0\bar{U}E'R/w^2$ which defines the distribution of contact pressure.

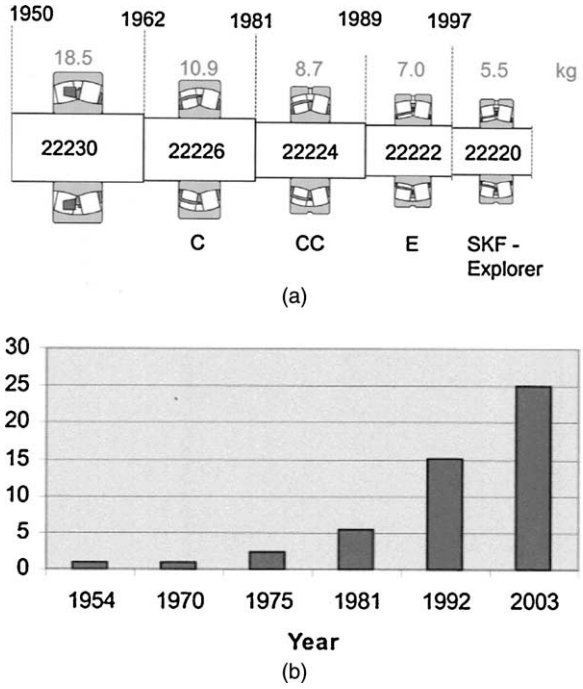

Fig. 9. The development of rolling element bearings over the last 50 years. (a) Each bearing has effectively the same load and speed rating while the mass and linear dimensions have reduced significantly. (b) The figures plotted represent the multiples of change in life relative to the corresponding bearing in 1954. The improvement in performance since 1970 is dramatic.

of more than three and contact stresses have typically doubled. Even more striking than this is the relative life rating shown in Fig. 9(b) which has increased by a factor more than 20 in the same period. Because bearing steels are now so much cleaner, some of the benefits of developing protective residual stresses in the outer layers of the races through a shakedown mechanism can be achieved; rolling element bearings are now much more likely to fail because of stress raisers introduced into the surface of the races by entrained particulate contaminants than because of subsurface material defects.

6.2. Cams and followers

To improve their thermodynamic efficiency and provided cleaner combustion modern gasoline engines have more poppet valves than would have been the case a few decades ago—typically four, or perhaps even five, per cylinder. This means that there has been a corresponding increase in the number of cams and followers. For the greater part of its operating cycle the cam/follower contact which is supplied with the engine lubricant is protected by a hydrodynamic, or elasto-hydrodynamic, oil film. However, as a consequence of its non-steady operating regime there are two points per cycle when the entraining velocity between the cam and follower surfaces necessarily falls to zero. To protect the surfaces around these points, when the ehl film is exceedingly thin, oil formulators include in their additive package an anti-wear agent to protect the integrity of surfaces—without this addition, almost universally a form of zinc dialkyldithiophosphate (ZDDP), the cam and follower surfaces are likely to fail prematurely by a characteristic process known as 'scuffing' or 'scoring'.

The total life of wearing components such as those in an ICE valve train can be thought of as having three stages. Initially, when two fresh surfaces on two new components are run together there will be a period of 'running-in', during which there will be some local plastic deformation of the higher asperities left by the manufacturing process on each of the surfaces. This usually results in a modest improvement, i.e. a reduction, in surface roughness. In a well-behaved system, this stage will be succeeded by an extended period of satisfactory operation during which there is virtually no change in the observed surface roughness values. Although the surfaces continue to wear, i.e. to lose material, it is at a very modest rate—wear rates of successfully operating contacts of this sort are extraordinary low when measured in terms of the often quoted specific wear rate coefficient; values of less than 10^{-9} mm^3 N^{-1} m^{-1} are typical. The nature of the wear process during this mild wear stage often appears to be a form of burnishing or very fine polishing with very fine score lines indicating a very small scale abrasive component. Eventually, either because of some change in environmental or operating conditions, the surfaces grow rougher, the wear rate begins to increase and if remedial action is not taken the surfaces scuff.

During the sliding process, if the contact pressure on an asperity of the softer surface is greater than the shakedown limit, then it will undergo some element of plastic deformation. There will be a reduction in height and an increase in radius and these changes will lead to the contact pressure dropping to just the numerical value allowed by shakedown. By applying this condition to each asperity, the new or modified surface profile of the softer surface could be obtained. Through this mechanism the separation of the surfaces decreases and wear 'flats' or 'plateaux' are effectively burnished onto the softer surface. The limit of this process occurs when all the softer asperities have become worn flat, and such a profile will maximise the shakedown load. Using this argument, for the cases of both spherical and cylindrical asperities, Kapoor et al. [27] have derived expressions for the limits of the elastic regime of operation. These can be usefully displayed on appropriate shakedown maps as the frontiers between the areas of stable operation and those in which incremental plastic flow would be expected for each repeating cycle. Such a map for a two-dimensional surface, i.e. one consisting of parallel cylindrical asperities, is illustrated in Fig. 10; the ordinate is a non-dimensional pressure $\bar{P}$ defined by the expression $\bar{P} = \{P_s/p_s\}/\{N\sqrt{R\sigma}\}$ where P_s is the shakedown load intensity, p_s is the material shakedown pressure and is related to its hardness H. R is the asperity radius of the harder surface, σ the r.m.s. roughness of the harder surface

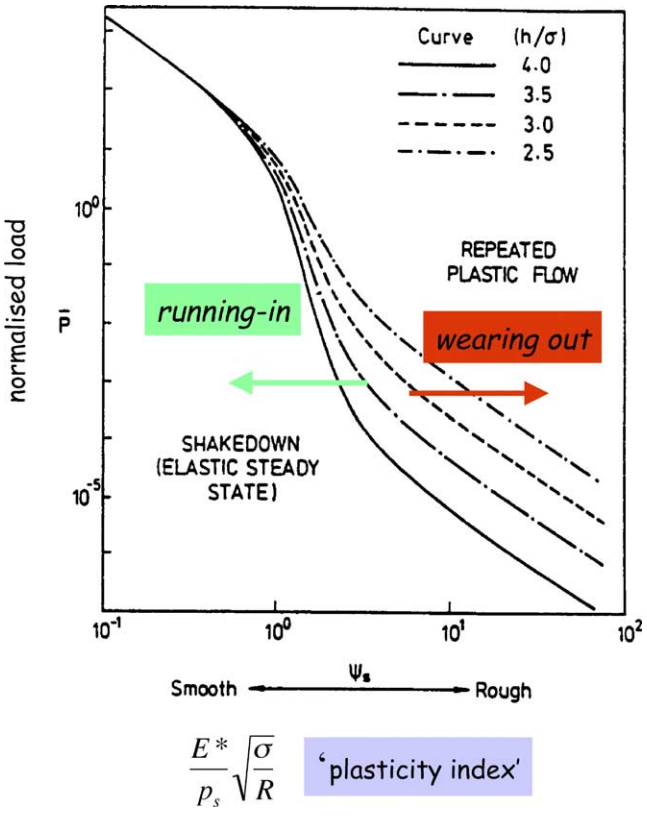

Fig. 10. A shakedown map for repeated line loading. The operational space described by the normalised load intensity $\bar{P}$ and the plasticity index in repeated sliding Ψ_s is divided into a benign shakedown region and a hazardous area where repeated plastic flow or ratchetting is to be expected with each load cycle, after Ref. [26].

and N the asperity density per unit length of the surface measured transversely to the 'lay'. The abscissa Ψ_s is termed the 'plasticity index for repeated sliding' defined as $\Psi_s = \{E^*/p_s\}\{\sigma/R\}^{1/2}$.

Numerical values of the plot of $\bar{P}$ vs. Ψ_s can be obtained for a Gaussian distribution of peak heights terminating at a chosen maximum peak height h expressed as a multiple of σ. Contacts whose operating values of $\bar{P}$ plot below the curves, i.e. in the bottom left hand portion of the map, would be expected to shakedown to an elastic state. Those in which the combination of material properties, topography and loading conditions lead to operating points above the curves will exhibit repeated plastic flow on each pass and would be expected to maintain a relatively high wear rate.

The process of running-in a tribological contact, even in the absence of any contribution through the mechanism of work-hardening, will lead to a reduction in the measure of roughness σ with an associated increase in the value of the asperity radius R and thus in the movement of the operating point on the map of Fig. 10 from right to left. Conversely, 'wearing-out', which involves some form of increase in surface roughness will move conditions from left to right. If these changes involve crossing the stability frontier then we should expect, in then first case, the initiation of a very stable low wear regime and, in the second, the loss of such a state and a rapidly accelerating wear rate leading to collapse by scuffing.

Experimental verification of this contribution to the mechanism of scuffing and, in particular, the part played by the anti-wear films generated as a result of the chemical activity of ZDDP have been studied by Bell [28] and Bell and Willemse [29] who carried our a series of fired engine tests, using different lubricants, in a test cycle that incorporated both hot and cold operating conditions. Detailed measurements were made to follow the evolution of surface roughness and wear. The tests were stopped at regular intervals to examine and measure the contacting surfaces of the cams and followers. Their observations demonstrated that the observed increase in surface roughness which preceded the initiation of scuffing was consistent with the operating point of the contact, plotted in the shakedown map, moving across the boundary from the region of elastic shakedown, and thus safe operation, into the region of repeated plastic flow where surface life would be expected to be much curtailed.

6.3. Wheel–rail contacts

The idea of using rails to both guide wheels and reduce rolling friction has a long history. The essential characteristic of the wheel–rail contact is its extreme vertical stiffness. The area of contact is small, and in addition to supporting the weight of the train, traction, braking and guidance forces must be carried across the interface. The consequential extremely high stresses have implications not only for the wheel and rail themselves but for their supporting structures in both the permanent way and the vehicle: these have been recently discussed by Smith [30]. In particular, increasing axle loads and higher traction stresses mean that in the case of many wheel rail contacts the protective mechanisms discussed in this paper may be insufficiently powerful to prevent the initiation of ratchetting behaviour in the upper-most material of the rail head as it experiences the loads generated by each passing wheelset. The ratchetting material accumulates strain up to its limiting ductility. Beyond this, material either detaches as wear debris or, if surrounded by less damaged material, continued load cycling can generate a network of crack-like flaws or weaknesses, typically initiating at a shallow angle of between 10 and 20° to the interface: these features continue to grow with each load cycle. Fluid—rainwater—which enters the cracks can assist this process either by reducing friction between the crack faces or by hydraulically transmitting the contact load as a tensile stress near the crack tip. If the wear rate is greater than the development of these fatigue cracks, sometimes known as 'gauge-corner cracking', then the deterioration of the rail is benign. However, if the wear rates are low then it is possible for the shallow cracks to bifurcate and turn into the body of the rail head so moving out of the region of surface residual

compressive stress into a deeper region where the residual stresses are tensile. If the branched crack turns back to the surface then a part of the rail surface detaches—a form of damage which is clearly visible on inspection. However, downward branches are very difficult to detect and if they eventually grow into the region of the rail which sees the gross longitudinal bending stresses can cause complete fracture of the rail.

The consequences of this sequence can be extremely serious. On 17 October 2000, a train travelling at 185 kph derailed at Hatfield just north of London, killing four passengers. The immediate cause was identified as a broken rail. This rail in question had, in fact, fragmented into more than 200 pieces. An examination of the UK network led to the discovery of more than 2000 sites containing potentially dangerous cracks. Severe speed restrictions were imposed and repair and replacement took many months. The UK rail system had been privatised in 1996 splitting the former nationalised British Rail into more than 125 companies and separating operations from infrastructure. As a consequence of the Hatfield accident, Railtrack, the infrastructure company went into receivership.

7. Conclusions

Shakedown theorems adopted from the theory of material plasticity can be used to establish rational design criteria for rolling and sliding tribological contacts thereby allowing significant increases in working loads or improvements in the efficient use of material. They do, however, expose the potential damaging effect of frictional tractions when sliding accompanies rolling. The statical theorem establish a safe limit which, if not exceeded, ensures that repeated plastic deformation will be avoided in the steady state, while the kinematic theorem establishes conditions in which incremental collapse or ratchetting of strain is to be expected. The results of the application of these ideas can be displayed in charts or shakedown maps on which the operating point of a tribological contact can be plotted and which delineate the various possible responses of the component. By making use of both an understanding of the generally beneficial and protective effects of the residual stresses developed in the shakedown process and improvements in materials integrity substantial gains in component performance can be achieved.

References

[1] Ham G, Rubin CA, Hahn GT, Bhargave V. Elastic–plastic finite element analysis of repeated two-dimensional rolling sliding contact. J Tribol 1988;110:44–9.

[2] Kulkarni S, Hahn GT, Rubin CA, Bhargava V. Elastic–plastic finite element analysis of three dimensional pure rolling contact above the shakedown limit. Trans ASME J Appl Mech 1991;58: 347–53.

[3] Van Dang K, Maitournam MH. Steady state flow in classical plasticity: applications to repeated sliding and rolling contact. J Mech Phys Solids 1993;41:1691–710.

[4] Yu MM-H, Moran B, Keer LM. A direct analysis of 3-D elastic–plastic rolling contact. ASME J Tribol 1995;117:234–43.

[5] Yu MM-H, Moran B, Keer LM. A direct method of 2-D elastic–plastic rolling contact. ASME J Tribol 1993;115:227–36.

[6] Yu MM-H, Moran B, Keer LM. A simplified direct method for cyclic strain calculation: repeated rolling sliding contact on a case-hardened half-plane. ASME J Tribol 1996;118:329–34.

[7] Sakae C, Keer LM. Application of direct method for a non-linear kinematic hardening material under rolling/sliding contact: constant ratchetting rate. J Mech Phys Solids 1997;45(9): 1577–94.

[8] Calladine CR. Plasticity for engineers. Oxford: Pergamon; 1969.

[9] Johnson W, Mellor PB. Engineering plasticity. London: Van Nostrand; 1973.

[10] Melan E. Der spannungsgudstand eines Henky-Mises schen kontinuums bei verlandicher bealstung. Sitzungberichte der Ak. Wissenschaften Wien 1938; Ser. 2A, 147:73.

[11] Koiter WT. A new general theorem on shakedown of elastic–plastic structures. K Ned Akad Wet 1956;B59:24–32.

[12] Kapoor A, Johnson KL. Effect of changes on contact geometry on shakedown of surfaces in rolling/sliding contact. Int J Mech Sci 1992; 34(3):223–39.

[13] Poritsky H. Stresses and deflections of cylindrical bodies in contact. ASME J Appl Mech 1950;17:191–201.

[14] Smith JO, Liu CK. Stresses due to tangential ands normal loads on an elastic solid with application to some contact stress problems. ASME J Appl Mech 1953;20:157–66.

[15] Sackfield A, Hills DA. Some useful results in the classical Hertz contact problem. J Strain Anal 1983;18:101–10.

[16] Johnson KL. A shakedown limit in rolling contact. Proceedings of fourth US national congress of applied mechanics, Berkeley, ASME.

[17] Johnson KL. Contact mechanics. Cambridge: Cambridge University Press; 1985.

[18] Johnson KL. Applied stress analysis. Hyde TH, Ollerton E, editors. Amsterdam: Elsevier.

[19] Johnson KL. The application of shakedown principles in rolling and sliding contacts. Eur J Mech A Solids 1992;11(Special issue): 155–72.

[20] Ponter ARS. A general shakedown theorem for elastic–plastic bodies with work-hardening. Third international conference on structural mechanics in reactor technology, London 1976.

[21] Kapoor A, Williams JA. Shakedown limits in rolling/sliding point contacts an anisotropic half-space. Wear 1994;191:256–60.

[22] Kapoor A, Williams JA. Shakedown limits on sliding contacts on a surface hardened half-space. Wear 1994;172:197–206.

[23] Kapoor A, Williams JA. The effect of interfacial shear strength on the performance of coated surfaces in repeated sliding. Trans ASME J Tribol 1996;119:541–8.

[24] Johnson KL, Jefferis JA. Plastic flow and residual stresses in rolling and sliding contact. Proceedings symposium on fatigue in rolling contact. London: Institution Mechanical Engineers; 1963 p. 54–65.

[25] Johnson KL. Correlation of theory and experiment in research on fatigue in rolling contact. Proceedings symposium on fatigue in rolling contact. London: Institution Mechanical Engineers; 1963 p. 155–9.

[26] Tallian TE. Failure atlas for hertz contact machine elements. New York: ASME; 1992.

[27] Kapoor A, Williams JA, Johnson KL. The steady-state sliding of rough surfaces. Wear 1994;175:81–92.

[28] Bell JC. Proc Inst Mech Eng Part J 1998;212:243–57.

[29] Bell JC, Willemse PJ. Inst Mech Eng Part J 1998;212:259–69.

[30] Smith RA. The wheel–rail interface—some recent accidents. Fatigue Fract Eng Mater Struct 2003;26:901–7.