EFFECT OF SOIL VARIABILITY ON THE BEARING CAPACITY OF FOOTINGS ON MULTI-LAYERED SOIL

By

Yien Lik Kuo

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY (PHD)

SCHOOL OF CIVIL, ENVIRONMENTAL AND MINING

ENGINEERING

OCTOBER 2008

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Foundations of engineering structures are designed to transfer and distribute their loading to the underlying soil and/or rock. This design is required to satisfy three main design criteria, namely the ultimate bearing capacity of the foundations (i.e. strength); the total and differential settlements (i.e. serviceability); and the economic feasibility of the foundation. This study focuses on the first of these criteria, that is the ultimate bearing capacity of shallow foundations and, in particular, foundations on cohesive and cohesive-frictional soil.

Bearing capacity failure occurs as the soil supporting the foundation fails in shear, which may involve either a general, local or punching shear failure mechanism (Bowles, 1988). For these different failure mechanisms, different methods of analysis are used. Estimation and prediction of the ultimate bearing capacity of a foundation is one of the most significant and complicated problems in geotechnical engineering (Poulos et al., 2001). Consequently, there is extensive literature detailing both theoretical and experimental studies associated with this issue. A list of the principal contributions to the study of this subject may be found, for example, in Terzaghi (1943), Hansen (1970), Vesic (1973), Chen and McCarron (1991) and Tani and Craig (1995). The focuses of these studies were on the estimation of the ultimate bearing capacity of the foundation under the combination of vertical, horizontal and moment loading, as well as the effect of foundation shape, soil rigidity, load inclination, tilt of the foundation base, ground surface inclination and the depth of the foundation on the ultimate bearing capacity of footing. For example, the ultimate bearing capacity equation for rectangular footings, which was suggested by Vesic (1973), is an extension of the theory first proposed by Terzaghi (1943).

In nature, soil deposits are often formed in discrete layers. As a result, footings are usually supported by multi-layered soil profiles, which influence the depth of the failure surface and the bearing capacity of the foundation. Traditionally, deterministic methods have been employed in practice to evaluate the ultimate bearing capacity of foundations on layered soil (e.g. Brown and Meyerhof, 1969; Meyerhof, 1974; Chen, 1975; Hanna and Meyerhof, 1980). The finite element method, which can handle very complex layer patterns, has also been applied to this problem and reliable

estimates have been presented (e.g. Griffiths, 1982a; Love et al., 1987; Burd and Frydman, 1997; Merrified et al., 1999). These studies are very useful although their applications are limited to a few situations. In many cases, the soil may be deposited in several layers. For such cases, reliable estimation of bearing capacity is extremely complicated. Modern computation techniques, such as the finite element method, require considerable effort before reliable estimates can ultimately be achieved. The question arises whether a simple hand calculation technique can be developed to provide reliable estimates of the ultimate bearing capacity of layered soils.

The deterministic methods are based on the assumptions of *uniformity*; that is, the properties of the soil in each layer are assumed to be uniform and, hence, homogeneous. However, natural soil deposits are inherently anisotropic due to the manner in which they are deposited, which is usually in horizontal layers. The soil properties are not distributed randomly, but in a semi-continuous fashion. It has been observed that the performance of foundations is considerably affected by the inherent spatial variability of the soil properties (Griffiths and Fenton, 2001). To date, some research has been undertaken investigating the probabilistic analysis of the settlement of foundations supported on single-layered soil profiles incorporating spatial variability (e.g. Griffiths and Fenton, 2001; Griffiths et al., 2002). The present study, however, focuses on the effects of a multi-layered, weightless and spatially random soil profile on the performance of foundations, which includes the bearing capacity of different footing sizes, as very limited work has been done in this field.

1.2 AIMS AND SCOPE OF THE STUDY

This research aims to investigate and quantify the effect of the spatial variation in layered soil on the ultimate bearing capacity of strip footings. This study also aims to develop meta-models to provide a first approximation of ultimate bearing capacity of strip footings on weightless multi-layered soil profile. A meta-model is described as a precise definition of the constructs and rules needed for creating semantic models. The meta-model used in the analysis is the artificial neural network. This investigation focuses on rough strip footings, which have a large footing length to footing width ratio, founded on two types of soils, namely purely cohesive soil and cohesive-frictional weightless soil.

In addition, this research seeks to provide:

- an in-sight into the effect of spatial variability of soil properties on the ultimate bearing capacity of strip footings founded on two-layered, purely cohesive weightless soil; and

- fast, simple and reliable meta-models to predict the ultimate bearing capacity of strip footings founded on purely cohesive, and cohesive-frictional, weightless multi-layered soil profiles.

It is noted that almost all previous research studies involving probabilistic analysis of footings on random soil are numerical, rather than experimental-based. This is because of the large variety of soil types and cases of multi-layered soil profiles that exist in nature. Establishing probabilistic analyses experimentally would be extremely ambitious and tedious, even if it is possible. Therefore, stochastic numerical modelling, combined with Monte Carlo simulation, has been adopted in this study. Random field simulation is adopted to simulate a series of plausible random soil profiles, taking into consideration the correlation of their properties. These issues are examined in detail later in the thesis. In addition, it should be noted that the study is concerned with the cohesion term of the bearing capacity problem. Accordingly, the soil is treated as a weightless soil.

1.3 THESIS OUTLINE

The primary focus of this thesis is to provide a rigorous study into the behaviour of strip footings, that is, those with infinite length over width ratio ($L/B = \infty$), on layered weightless soil. As mentioned in the previous section, most engineering properties of soil exhibit natural variability in a spatial extent. Consequently, this study incorporates random field theory, which is capable of quantifying and simulating the spatial variability of natural soil, and is facilitated, in particular, by finite element analysis. Incorporating random field simulation into finite element analysis allows the extent to which spatial variability affects the bearing capacity and load-settlement

behaviour of the footing to be evaluated using Monte Carlo simulation. This implementation of random field theory is addressed in Chapter 3.

A brief review of research into the bearing capacity and load-settlement response of footings on layered soil will be presented in Chapter 2. While very few numerical studies on the bearing capacity of footings have addressed the influences of spatial variability in single-layered soil, none exists for multi-layered, spatially random soil profiles. This will become more apparent in Chapter 2.

Chapter 3 provides background to selected aspects of classical plasticity and random field theories, as well as artificial neural networks. It also includes detailed discussion on the numerical formulations adopted in this study. In Chapter 4, greater detail is given regarding the process adopted for studying footings using these numerical formulations. This includes a discussion of finite element mesh arrangements, footing interfaces, implementation of random field simulation and cross-validation.

As an overview, the research presented in this thesis can be divided into three principal areas:

(1) the investigation of bearing capacity and load-displacement response of rough strip footings on two-layered, purely cohesive, spatially random soil;

(2) the development of a bearing capacity meta-model for rough strip footings on a multi-layered, purely-cohesive, homogeneous soil profile;

(3) the development of a bearing capacity meta-model for rough strip footing on a multi-layered, cohesive and frictional, homogeneous soil profile.

The structure of the thesis reflects the three main topics listed above. In Chapters 5 to 7, the results obtained from the numerical studies are presented. A wide range of cases and problems, which are separated according to soil types are examined and the results are discussed.

A summary and conclusion of this study are presented in Chapter 8, including recommendations for future research.

CHAPTER 2

HISTORICAL REVIEW

2.1 INTRODUCTION

In order to provide a general background to subsequent Chapters, a summary of previous research into the bearing capacity of footings on multi-layered and spatially random soil profiles is presented.

Foundation designs must satisfy both strength and serviceability criteria. The soil beneath the foundations must be capable of carrying the structural loads placed upon it without shear failure and consequent settlements being tolerated for the structure it is supporting. Rupture surfaces are formed in the soil mass upon exceeding a certain stress condition. Hence, bearing capacity is defined as the capacity of the underlying soil and footing to support the loads applied to the ground without undergoing shear failure and without accompanying large settlements (Das, 1997).

2.2 BEARING CAPACITY OF FOOTING

Shallow footings, which are often referred to as foundations, can be classified into two general categories (1) those that support a single structural member, usually referred to as “spread or pad footings”, and (2) those that support two or more structural members, referred as “combined or strip footings”, as shown in Figure 2.1.

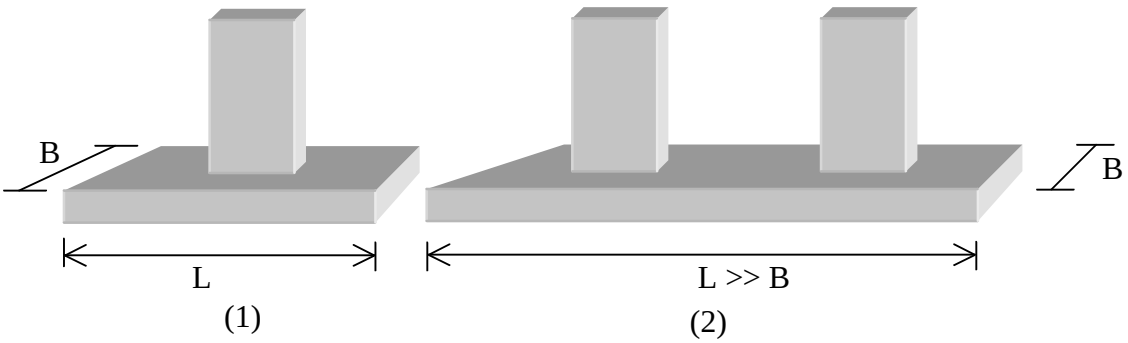


Figure 2.1 Foundation types: (1) spread or pad footing (2) combined or strip footing.

The foundations transmit loads from a structure to the underlying soil. Their load is supported by the shear strength and dead weight of the underlying and surrounding

soil. The bearing capacity of a foundation is defined as the load per unit area of the foundation at which the soil shear failure occurs (Terzaghi, 1943).

Consider a footing resting on the surface of clay or sand. As the foundation load gradually increases, the settlement of the foundation also increases. At a certain point, that is when the load per unit area is equal to the bearing capacity of the footing, a sudden failure in the soil will generally take place. Such failure is often accompanied by a large increment of footing settlement and the failure surface may extend to the ground surface (Das, 1997). Figure 2.2 shows a classic example of bearing capacity failure, which occurred to the Transcona Grain Elevator, Canada, in 1913.

NOTE:
This figure is included on page 9 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.2 Bearing capacity and excessive settlement failure of Transcona Grain Elevator, Canada. (*After Baracos, 1957*)

The bearing capacity failure of a strip footing on a single-layered, homogeneous soil is presented in Figure 2.3. In nature, soils often consist of discrete layers, and, as a

result, foundations for engineering structures are usually founded on multi-layered soil profiles. If a footing is placed on the surface of a layered soil and the thickness of the top layer is large compared with the width of the footing, then the bearing capacity of the soil and the displacement behaviour of the footing can be estimated, to sufficient accuracy, using only the properties of the upper layer (Poulos et al., 2001).

Figure 2.3 Bearing capacity of footing on single homogeneous soil (*After Das, 1997*).

However, if the thickness of the uppermost layer is equal to or less than the width of the foundation, such an approach introduces significant inaccuracies and is no longer appropriate (Poulos et al., 2001). This is because the zone of influence of the footing, including the potential failure zone, may extend to a significant depth, and thus two or more layers within that depth range will affect the bearing behaviour of the footing.

Consider a foundation supported on a two-layered soil profile. If the depth of the top layer, *H*, is relatively small compared to the foundation width, *B*, punching shear failure will occur in the upper soil layer followed by a general shear failure in the lower soil layer. Such bearing capacity failure of a footing on a two-layered soil, is illustrated in Figure 2.4.

In general, shear failure of a soil mass supporting a foundation will occur in one of three modes; namely, general shear, local shear and punching shear (Das, 1997). General shear failure is shown in Figure 2.5, and can be described as follows: the soil wedge immediately beneath the footing (an active Rankine zone acting as part of the footing) pushes Zone II laterally. This horizontal displacement of Zone II causes Zone III (a passive Rankine zone) to move upward. Although bulging of the ground surface may be observed on the both sides of the footing at a stress level below failure, failure

usually occurs on only one side of the footing (Das, 1997). For example, an isolated structure may tilt substantially or completely overturn, as shown previously in Figure 2.2. A footing restrained from rotation by the structure will increase structural moments which may lead to collapse or excessive settlement.

NOTE:
This figure is included on page 11 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.4 Bearing capacity of footing on two-layered soil (A strong layer overlying a soft layer soil). (*After Das, 1997*)

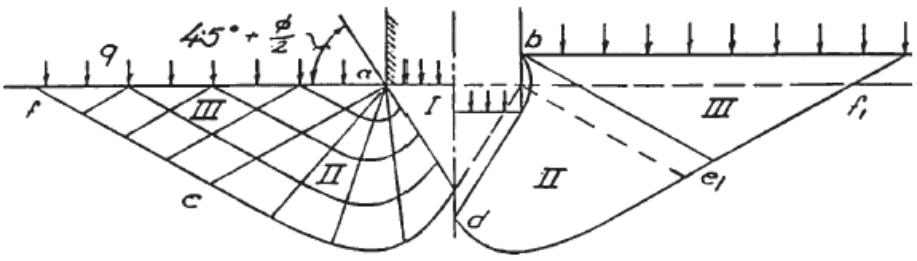


Figure 2.5 General shear failure concept. (*After Vesić, 1973; Das, 1997*)

The variation of foundation settlement and load per unit area, q, is also shown in Figure 2.6. In the initial loading phase, the settlement increases as the load per unit area increases. At a certain point, i.e. when load per unit area equals q_u, a sudden

failure in the soil supporting the foundation will take place. This load per unit area is usually referred to ultimate bearing capacity of the foundation.

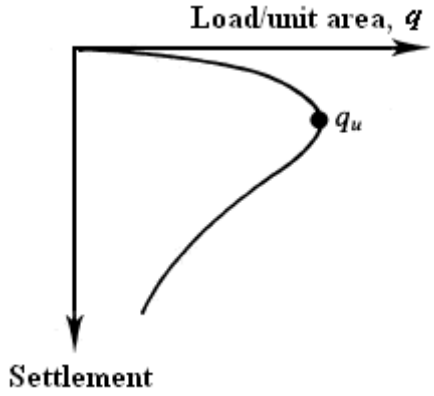


Figure 2.6 Load settlement plot for general shear failure type. (*After Vesić, 1973; Das, 1997*)

Punching shear failure presents little, if any, ground surface evidence of failure, since the failure occurs primarily in soil compression immediately beneath the footing (as illustrated in Figure 2.7). Footing stability (i.e. no rotation) is usually maintained throughout failure (Das, 1997). The failure surface in the soil will not extend to the ground surface. Beyond the ultimate failure load, q_u, the load settlement plot will be practically linear, as illustrate in Figure 2.7.

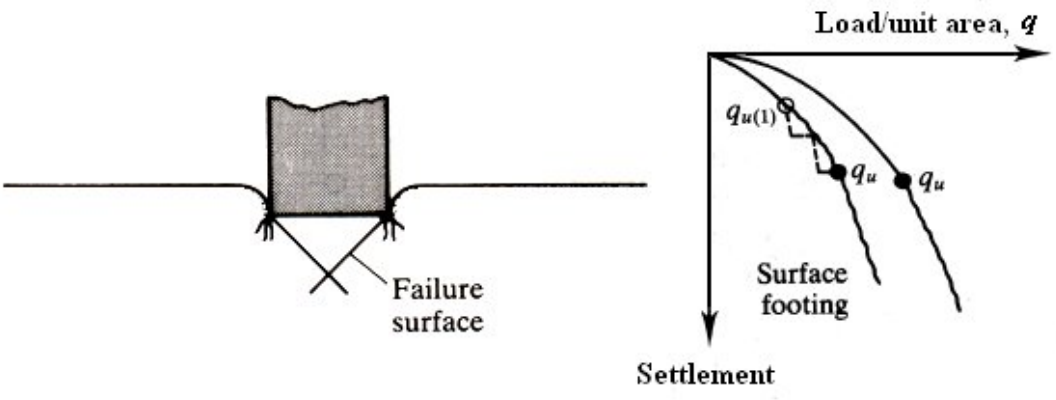


Figure 2.7 Punching shear failure and its load settlement plot. (*After Vesić, 1973; Das, 1997*)

Local shear failure (shown in Figure 2.8) may exhibit both general and punching shear characteristics, soil compression beneath the footing, and possible ground

surface bulging (Das, 1997). As the load per unit area on the foundation equals $q_{u(1)}$ (refer to load settlement plot in Figure 2.8), the foundation movement will be accompanied by sudden jerks. The $q_{u(1)}$ is referred to as first failure load (Vesić, 1973) and it is less than the ultimate failure load per area, q_u. Therefore, this type of failure is referred to as local shear failure in soil.

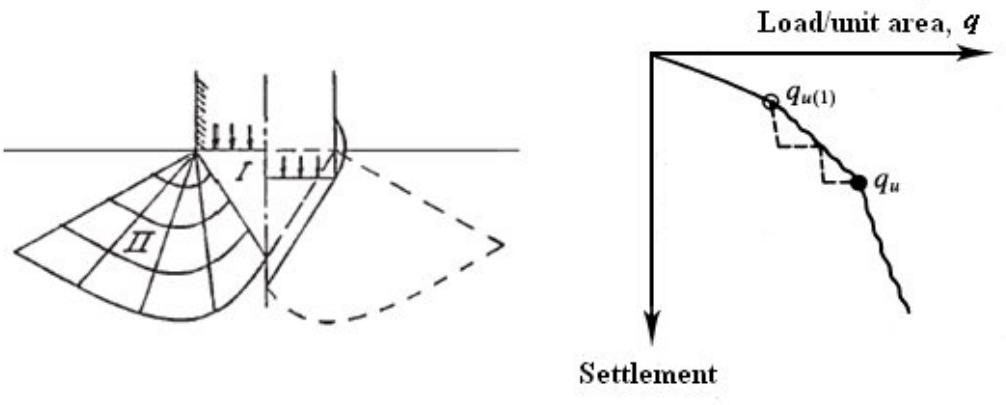


Figure 2.8 Local shear failure and its load settlement plot. (*After Vesić, 1973; Das, 1997*)

In practice, the general shear failure case is the one normally analysed, as the other failure modes are accounted for implicitly in settlement calculations (Coduto, 2001). A rational approach to predict the bearing capacity of a foundation was developed by Terzaghi (1943) and is given as follow:

For a square footing:

$$q_{ult} = 1.3cN_c + \sigma_z N_q + 0.4\gamma B N_\gamma \tag{2.1}$$

For a strip footing:

$$q_{ult} = cN_c + \sigma_z N_q + 0.5\gamma B N_\gamma \tag{2.2}$$

where c is the soil cohesion; σ_z is the vertical stress at the base of the foundation; γ is the unit weight of soil; B is the width of the foundation; N_c, N_q and N_γ are non-

dimensional bearing capacity factors, which can be derived from following relationships:

$$N_q = e^{\pi \tan\phi} \tan^2\left(45 + \frac{\phi}{2}\right) \tag{2.3}$$

$$N_c = \frac{N_q - 1}{\tan\phi} \tag{2.4}$$

where ϕ is the internal friction angle of the soil.

However, significant discrepancies have been noted in the values proposed for N_γ (Poulos et al., 2001). It is has not been possible to obtain a rigorous closed form expression for N_γ, but several authors have proposed the following approximations:

$$N_\gamma \approx \begin{cases} 1.8(N_q - 1)\tan\phi & \text{(Hansen, 1970)} \\ (N_q - 1)\tan(1.4\phi) & \text{(Meyerhof, 1963)} \\ 0.0663e^{9.3\phi} \text{ smooth} & \text{(Davis and Booker, 1971)} \\ 0.1054e^{9.6\phi} \text{ rough} & \text{(Davis and Booker, 1971)} \\ 2(N_q + 1)\tan\phi & \text{(Vesic, 1975)} \end{cases} \tag{2.5}$$

More recently, quasi-exact values for N_γ were obtained and presented by Hjiaj et al. (2005) and Makrodimopoulos and Martin (2005) using numerical finite element limit analysis. Davis and Booker (1973) compared the rigorous solutions, which were obtained using the theory of plasticity for a rigid plastic body, with the approximate values suggested by Terzaghi (1943), and the comparison clearly indicated that Terzaghi's approximation can overestimate the bearing capacity by factors as large as 3.

Although Equations 2.1 and 2.2, provide reasonable estimates of bearing capacity for general engineering purposes, recent work by researchers working in experimental soil mechanics has shown that the soil behaviour is highly non-linear (e.g. Chen and Mizuno, 1990; Head, 1997). Therefore, the simple assumption of superimposing Equations 2.1 and 2.2 do not necessarily hold (Davis and Booker, 1971).

The use of superposition simplifies the mathematical analysis considerably and, because of the complication that are introduced by the inclusion of self-weight, the general bearing capacity problem is usually solved in two stages. In the first stage, the analytical solution of Prandtl (1921), which assumes a weightless material, is used to give the bearing capacity factors N_c and N_q in closed form. In the second stage, the contribution of the soil weight is typically found using a numerical solution technique to give the bearing capacity factor N_γ.

While the exact values for N_γ remain unknown, the values for N_c and N_q, as given by Prandtl (1921), and Reissner (1924) are exact for a strip footing on a weightless soil. As discussed by Chen (1975), the analysis of cohesionless soil with self-weight is complicated by the fact that the shear strength increases with depth from a value of zero at the ground surface. This means that the Prandtl failure mechanism is no longer capable of yielding exact results, since any velocity discontinuity that is initially straight for the weightless case now becomes curved. This leads to the conclusion that the bearing capacity obtained using this mechanism can, at best, only be an upper bound on the correct value. Similar conclusions may be drawn regarding the mechanism suggested by Hill (1949).

As mentioned earlier, due to the complexities that are associated with the introduction of self-weight, this study is focus on the cohesion term of the bearing capacity problem. Accordingly, it is assumed that the soil is weightless.

2.3 MULTI-LAYERED SOIL PROFILE

Due to the sedimentation and weathering processes, most natural occurring soils are formed in discrete layers; thus, in most cases, footings are founded on layered soil. Bowles (1996) pointed out that there are three general cases of footings on a two-layered soil, as follows:

Case 1. Footing on layered clays.

a. Upper layer weaker than underlying layer ($c_1 > c_2$)

b. Upper layer stiffer than underlying layer ($c_1 < c_2$)

Case 2. Footing on layered c-ϕ soils with a, b same as Case 1.

Case 3. Footing on layered sand and clay soil.

a. Upper sand layer overlying clay layer

b. Upper clay layer overlying sand layer

In many cases, soils may be deposited in several layers and the zonal boundary along the contact between two geological formations is often indistinct.

In order to simplify the problem, consider a strip footing of width B, founded on the surface of soil deposit consisting of three different horizontal layers, as illustrated in Figure 2.9. The thickness of the upper two layers is assumed to be less than B. The strength of each layer is characterised by the Mohr-Coulomb shear strength parameters (c and ϕ), and its self-weight, γ. Various cases of possible practical problems are outlined as follows and are indicated in Table 2.1.

1. Layered clays with a soft clay layer at the centre (Case 4)
2. Layered clays with a stiff clay layer at the centre (Case 5)
3. Three clay layers, strengthening with depth (Case 6)
4. Three clay layers, weakening with depth (Case 7)
5. A layer of sand at the centre of two clay layers (Case 8)

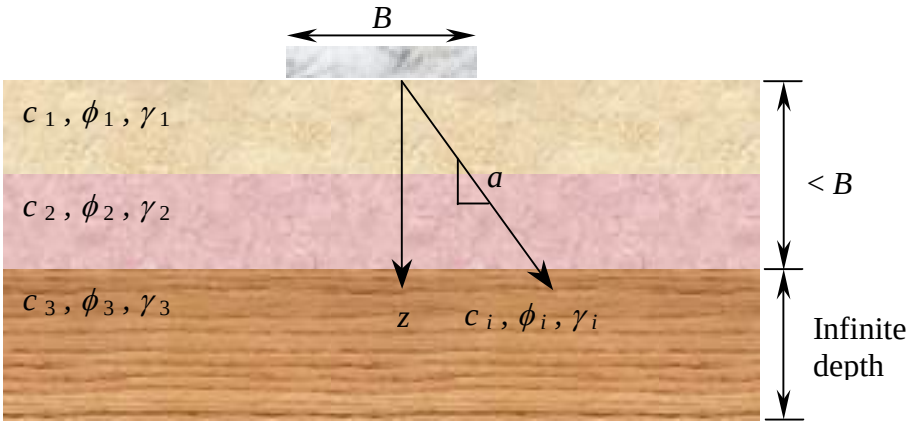


Figure 2.9 Strip footing on three-layered soil deposit.

6. Layered sands with a soft sand layer at the centre (Case 9)
7. Layered sands with a stiff sand layer at the centre (Case 10)
8. Three sand layers, strengthening with depth (Case 11)

9. Three sand layers, weakening with depth (Case 12)
10. A layer of clay at the centre of two sand layers (Case 13)

Table 2.1 General cases for soil deposits with three layers.

Case	Layer 1			Layer 2			Layer 3		
	c_1	ϕ_1	γ_1	c_2	ϕ_2	γ_2	c_3	ϕ_3	γ_3
4	c_1	0	0	$< c_1$	0	0	$= c_1$	0	0
5	c_1	0	0	$> c_1$	0	0	$= c_1$	0	0
6	c_1	0	0	$= az_1\, c_1$	0	0	$= az_2\, c_1$	0	0
7	c_1	0	0	$= \{1/(az_1)\}\, c_1$	0	0	$= \{1/(az_2)\}\, c_1$	0	0
8	c_1	0	γ_1	0	ϕ_2	$= \gamma_1$	$= c_1$	0	$= \gamma_1$
9	0	ϕ_1	γ_1	0	$< \phi_1$	γ_2	0	$= \phi_1$	$= \gamma_1$
10	0	ϕ_1	γ_1	0	$> \phi_1$	γ_2	0	$= \phi_1$	$= \gamma_1$
11	0	ϕ_1	γ_1	0	$= az_1\, \phi_1$	γ_2	0	$= az_2\, \phi_1$	$= \gamma_1$
12	0	ϕ_1	γ_1	0	$= \{1/(az_1)\}\, \phi_1$	γ_2	0	$= \{1/(az_2)\}\, \phi_1$	$= \gamma_1$
13	0	ϕ_1	γ_1	c_2	0	$= \gamma_1$	0	$= \phi_1$	$= \gamma_1$

Note: The parameter a is defined as the variation of strength with regard to depth, z, as shown in Figure 2.9.

In an assessment of bearing capacity and load-displacement behaviour of footings, if the thickness of the uppermost layer is significantly larger than the footing width, then a reasonably accurate estimation of the collapse load and displacement behaviour can be made by using the properties of uppermost layer alone (Poulos et al., 2001). However, if the proposed footing has large physical dimensions (e.g. offshore foundations) and the thickness of upper layer is comparable to the footing width, special consideration needs to be given to account for the effect of any soil layers within the zone of influence of the footing. For such cases, reliable estimation of the bearing capacity is more complicated. With aid of modern computation techniques, such as finite element method, reliable estimations can be ultimately achieved. However, these methods often require some experience and considerable effort. The question arises as to whether simple hand techniques can be devised to provide realistic first estimates of the ultimate bearing capacity of layered soil.

The following sections provide a brief review of previous research into this problem. Previous investigations into the behaviour of the bearing capacity of footings on layered soil takes one of two forms, namely experimental and numerical/theoretical studies. In the interest of brevity, a selective summary of this research, which has the greatest relevance to the thesis, is presented rather than a complete bibliography of all such investigations. Where possible, the review is given in chronological order.

2.4 PREVIOUS THEORECTICAL ANALYSES OF BEARING CAPACITY OF FOOTINGS ON A MULTI-LAYERED SOIL PROFILE

Until recently, most of the theoretical investigations into the bearing capacity problem have been limited to plane-strain conditions, as illustrated in Figure 2.10. In this problem, a strip footing of width B rests upon an upper layer of clay with undrained shear strength, c_{u1}, and thickness, H. This layer is underlain by another clay layer of undrained shear strength, $c_{u2} \neq c_{u1}$, and infinite thickness. The bearing capacity of this problem will be a function of two ratios, H / B and c_{u1} / c_{u2}. As shown in Figure 2.10, it is important to note that both clay soils are assumed to be weightless.

The bearing capacity of a strip footing on a single clay layer in the absence of surcharge is usually written in the form:

$$q_{ult} = c_u N_c \tag{2.6}$$

where, as indicated earlier, q_{ult} is the ultimate bearing capacity of footing, c_u is undrained cohesion and N_c is the bearing capacity factor. For the case of a two-layered soil profile, it is convenient to rewrite Equation 2.6 in the form (Merifield et al., 1999; Poulos et al., 2001):

$$N^*_c = \frac{q_{ult}}{c_{u1}} \tag{2.7}$$

where N^*_c is a modified bearing capacity factor which is a function of both parameters H / B and c_{u1} / c_{u2}. For a homogeneous profile, where $c_{u1} = c_{u2}$, then N^*_c is equal to the well-known Prandtl solution of $(2+\pi) = 5.14$ (Prandtl, 1921).

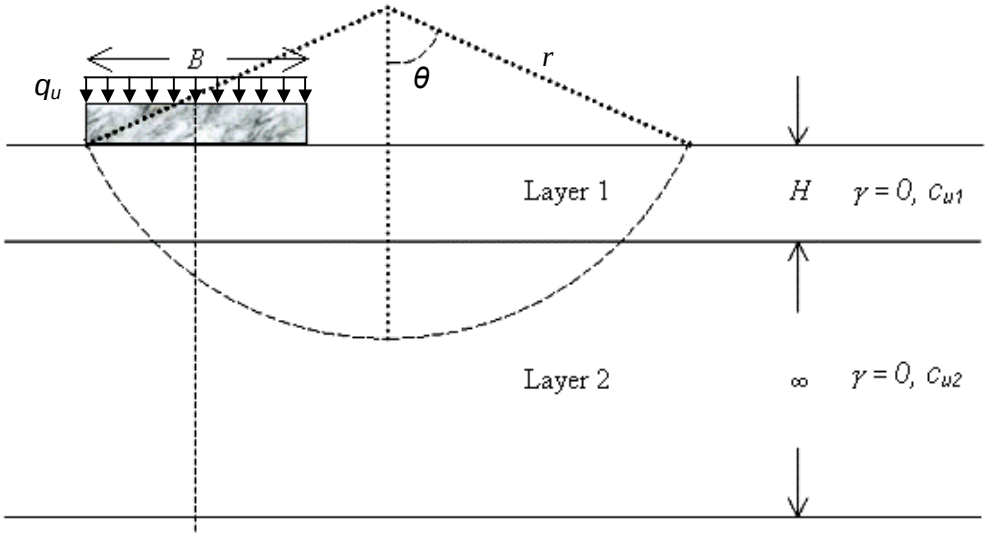


Figure 2.10 Bearing capacity approximation on two-layered clay profile. (*After Chen, 1975*)

To solve the problem of the bearing capacity of a shallow foundation on a two-layered clay profile, the limit equilibrium method can be adopted. The use of such a method in determining bearing capacities is widely accepted practice in geomechanics. In limit equilibrium analysis, the assumption of a simple semi-circular failure surface is usually made and simple static equations are applied by equating the applied and the resultant forces, thus enabling the bearing capacity to be subsequently determined (1975). The first to do this was Button (1953), followed by Chen (1975), both of whom used a circular arc to determine the upper bound value for the bearing capacity factor, N^*_c, for a strip footing, which was supported on a two-layered clay profile (as illustrated in Figure 2.10). Reddy and Srinivasan (1967) used a similar failure mechanism to obtain results by applying the limit equilibrium method.

For footings that are supported on stiff sand layers overlaying soft clay, one widely used approach to estimate bearing capacity is to assume that the sand layer spreads the

load from the footing to the underlain clay layer through a load spread mechanism, and the rupture surfaces are formed within the clay layer (Houlsby et al., 1989). The load in the sand layer is assumed to spread at a load-spread angle, β, so that the effective width of the loaded area, B', on the clay is $B' = B + 2D \tan \beta$, as illustrated in Figure 2.11 (Houlsby et al., 1989). In practice, the value of $\beta = \tan^{-1} 0.5$ is often adopted (Houlsby et al., 1989), although it is known that β is dependent on the shear strength of the sand (Burd and Frydman, 1997). Furthermore, an investigation by Brocklehurst (1993) found that β is also strongly influenced by the strength of the underlying clay. More recently, Burd and Frydman (1997) provided a series of figures of their investigation on the variation of β with respect to the shear strength ratio ($c_u / \gamma D$) and friction angle of sand using a numerical approach.

Several investigators have adopted kinematic analysis methods to obtain approximate solutions to the bearing capacity of a footing supported on sand overlying clay. Florkiewicz (1989), for example, presented solutions for a range of cases involving both cohesive and cohesionless soil. Michalowski and Shi (1995) used a similar approach to carry out a parametric study and the results were presented in the form of design charts. Their results were studied by Burd and Frydman (1996) and concluded that they may overestimate the bearing capacity due to the assumption that the sand is a fully associated material.

Figure 2.11 Load spread mechanism. (*After Houlsby et al., 1989*)

Burd and Frydman (1997) employed two different numerical methods, namely the finite element and finite difference methods, to conduct a parametric study on the bearing capacity of a sand layer overlying clay under plane-strain loading. They adopted unstructured meshes of six-noded triangular elements with three Gauss points. The tangent stiffness technique was used to solve the finite element algebraics. Most of the results were presented in figures in the form of plots illustrating the bearing capacity (p / γB) and shear strength ratios (c_u / γD) with respect to variation of ϕ and B/D. They also presented the results of their finding into the variation of load spread angle with different values of ϕ and c_u / γD and B/D.

Okamura et al. (1998) proposed an alternative bearing capacity formulation for thin sand overlying soft clay based on their new limit equilibrium mechanism, which illustrated in Figure 2.12. Consideration of equilibrium of forces acting on the sand block, including its self-weight, provides the following bearing capacity formulae for a strip footing:

$$q_{ult} = \left(1+2\frac{H}{B}\tan\alpha_c\right)\left(s_u N_c + p'_o+\gamma' H\right)+\left(\frac{K_p \sin(\phi'-\alpha_c)}{\cos\phi'\cos\alpha_c}\right)\left(\frac{H}{B}\right)\times \left(p'_o+\gamma' H\right)-\gamma' H\left(1+\frac{H}{B}\tan\alpha_c\right) \tag{2.8}$$

where α_c is an angle which is a function of the friction angle of the sand, ϕ, the geometries of the layer and footing, and the undrained shear strength of clay, s_u, such that:

$$\alpha_c = \tan^{-1}\left(\frac{\left(\sigma_{mc}/s_u\right)-\left(\sigma_{ms}/s_u\right)\left(1+\sin^2\phi'\right)}{\cos\phi'\sin\phi'\left(\sigma_{ms}/s_u\right)+1}\right) \tag{2.9}$$

where:

$$\sigma_{mc}/s_u = N_c s_u\left(1+\frac{1}{\lambda_c}\frac{H}{B}+\frac{\lambda_p}{\lambda_c}\right) \tag{2.10}$$

$$\sigma_{ms} / s_u = \frac{\sigma_{mc} / s_u - \sqrt{(\sigma_{mc} / s_u)^2 - \cos^2 \phi' \left((\sigma_{mc} / s_u)^2 + 1 \right)}}{\cos^2 \phi'} \qquad (2.11)$$

Figure 2.12 Failure mechanism proposed by Okamura et al. (1998) for thin sand overlying soft clay.

The parameters λ_p and λ_c are the normalised overburden pressure and the normalised bearing capacity of the underlying clay respectively, given by:

$$\lambda_p = \frac{p'_o}{\gamma' B} \quad \text{and} \quad \lambda_c = \frac{s_u N_c}{\gamma' B} \qquad (2.12)$$

Finally, the parameter K_p is the Rankine's passive earth pressure coefficient for the sand, which can be expressed as:

$$K_p = (1 + \sin\phi) / (1 - \sin\phi) \qquad (2.13)$$

Okamura et al. (1998) concluded that the Equation 2.8 provides reliable estimate of ultimate bearing capacity of footing on thin sand layer overlying clay for cases where $\lambda_c < 26$ and $\lambda_p < 4.8$, however, it is important to recognise that the method may not be applicable over full range of these value.

Many of the previous attempts to develop numerical models of the behaviour of footing on layered soils (e.g., Love et al., 1987; Griffiths, 1982a; Brocklehurst, 1993) were based on the use of a finite element method (FEM). It is well known that the finite element analysis method tends to overestimate the bearing capacity due to the addition of kinematic constraints imposed on the system by the specified volumetric strains associated with plastic flow (Sloan, 1981; Sloan and Randolph, 1982). Moreover, finite element analyses involving bearing capacity of frictional materials, particularly those with a large friction angle, are often have instability problems (e.g. Griffiths 1982b). This problem may be overcome by selecting 15-noded triangular elements, or by reduced integration procedures (Nagtegaal et al., 1974).

Recently, Merifield et al. (1999) presented the results of their investigation into the upper and lower bound solution of the bearing capacity factor of two-layered clay profile under strip footings by employing the finite element in conjunction with the limit theorems of classic plasticity. In their paper, the results were presented for different cases of H/B and c_1/c_2. The results of their extensive parametric study have been presented in both graphical and tabular form. Some of the results are presented in Figure 2.13 and 2.14.

More recently, Shiau et al. (2003) applied the finite element limit analysis to determine the bearing capacity of dense sand underlain by a soft clay layer. Bearing capacity of this problem is expressed in the non-dimensional form:

$$\frac{p}{\gamma B} = f\left(\frac{D}{B}, \frac{c_u}{\gamma B}, \frac{q}{B}, \phi'\right) \tag{2.14}$$

NOTE:
This figure is included on page 23 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.13 Bearing capacity factors, N^*_c, for two layered clay (H/B = 0.125 and H/B = 0.25). (*After Merifield et al., 1999*)

NOTE:
This figure is included on page 24 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.14 Bearing capacity factors, N^*_c, for two layered clay (H/B = 0.375 and H/B = 0.5). (*After Merifield et al., 1999*)

In their paper, they introduced a fan mesh for a lower bound analysis, and the fan of elements centred on the footing edge, where there is an abrupt change in the boundary conditions. The advantages of these stress fans are that each discontinuity allows a jump in the tangential stress, thus providing the potential for a rapid change in the stress field and a higher lower bound. The study covered a range of parameters, including the depth of sand layer (D/B), the friction angle of sand (ϕ'), the undrained shear strength of the clay ($c_u / \gamma B$), and the effect of a surcharge ($q / \gamma B$). The influence of footing roughness and increasing strength with depth for clay were also investigated. Some of the results are reproduced and presented in Figure 2.15.

Whilst, as shown above, research has been carried out with respect to profiles consisting of 2 soil layers, virtually no work has been carried out on calculating the bearing capacity associated with a number of thin layers. Bowles (1988) suggested a possible alternative for c - ϕ soils with n thin layers is to use average values of c and ϕ in the traditional bearing capacity equations (i.e. those of Terzaghi (1943), Meyerhof (1963), Brinch Hansen (1970) and Vesić (1973)) obtained by:

$$c_{av} = \frac{c_1 H_1 + c_2 H_2 + c_3 H_3 + \cdots + c_n H_n}{\sum H_i} \tag{2.15}$$

$$\phi_{av} = \tan^{-1} \frac{H_1 \tan\phi_1 + H_2 \tan\phi_2 + H_3 \tan\phi_3 + \cdots + H_n \tan\phi_n}{\sum H_i} \tag{2.16}$$

where c_i = cohesion of the layer of thickness H_i and ϕ_i = internal friction angle in of the i^{th} layer. This method assumes these average values of c and ϕ represent the average response of a multi-layered soil mass. As will be demonstrated later, Bowles' approach is a gross simplification of the bearing capacity problem.

NOTE:
This figure is included on page 25 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.15 Bearing capacity of sand on clay for (a) D/B = 0.25 (b) D/B = 1. (*After Shiau et al., 2003*)

2.5 PREVIOUS EXPERIMENTAL WORK

There has been little experimental verification of any of the numerical methods described above except by using model footings. Models of B = 25 to 75 mm × L = 25 to 200 mm are popular because the ultimate load can be developed in a small box of prepared soil in the laboratory using commonly available equipment of the order of 400 kN capacity. A full-size footing as small as 1 m × 1 m can develop ultimate loads of 3,000 to 4,000 kN so that very expensive site preparation and equipment are necessary to develop and measure loads of this magnitude. Moreover, the results produced from the experiments may be biased, because models, particularly on sand, do not produce reliable test results compared to full-scale prototypes due to scale effects.

Examples of studies that were based on a series of model tests are Brown and Meyerhof (1969) and Meyerhof and Hanna (1978). With a combination of theoretical analyses and model tests, Brown and Meyerhof (1969) provided approximate solutions for N_c of several cases and footing types, as follow:

For strip footing:

$$N^*_c = 1.5\left(\frac{H}{B}\right) + 5.14\left(\frac{c_2}{c_1}\right) \le 5.14 \qquad \text{for } c_2/c_1 \le 1 \tag{2.17}$$

For circular footing:

$$N^*_c = 1.5\left(\frac{H}{B}\right) + 6.05\left(\frac{c_2}{c_1}\right) \le 6.05 \qquad \text{for } c_2/c_1 \le 1 \tag{2.18}$$

where B is the diameter of the circular footing.

Perhaps the most widely used semi-empirical methods in practice are those provided by Meyerhof (1974) and Meyerhof and Hanna (1978). They suggested that, if the thickness of uppermost layer, H, is relatively small compared to the foundation width, B, punching shear failure will occur in the uppermost layer, followed by general shear failure in the lower soil layer, as illustrated in Figure 2.16.

With respect to this figure, the bearing capacity of a footing supported on layered soil can be expressed as (Meyerhof, 1974):

$$q_u = q_b + \frac{2(C_a + P_p \sin\delta)}{B} - \gamma_1 H \tag{2.19}$$

where δ is the angle of inclination of the passive force P_p with horizontal; q_b is the bearing capacity of bottom soil layer, which can be determined using

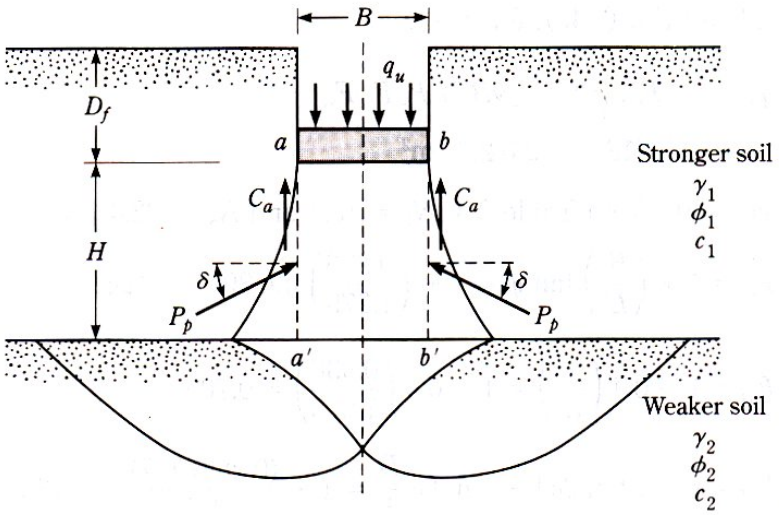


Figure 2.16 Punching shear models on layered soil (*After* Meyerhof, 1974; *Das, 1997*).

$$q_b = c_2 N_c + \gamma_1 (D_f + H) N_q + \frac{1}{2} \gamma_2 B N_\gamma \qquad (2.20)$$

and P_p is passive force per unit length of the faces *aa'* and *bb'* that is expressed in form

$$P_P = \frac{1}{2}\gamma_1 H^2 \left(1 + \frac{2D_f}{H}\right) \frac{K_{ph}}{\cos\delta} \qquad (2.21)$$

where K_{ph} is horizontal component of passive earth pressure coefficient. Let K_s is the punching shear coefficient and is presented in form (Hanna and Meyerhof, 1980)

$$K_s = K_{ph} \tan\delta / \tan\phi_1 \qquad (2.22)$$

Note $C_a = c_a H$ and therefore the Equation 2.19 can be rewritten as (Das, 1997):

$$q_u = q_b + \frac{2c_a H}{B} + \gamma_1 H^2 \left(1 + \frac{2D_f}{H}\right) \frac{K_s \tan\phi_1}{B} - \gamma_1 H \qquad (2.23)$$

The relationship between K_s and ϕ_1 is plotted and presented by Hanna and Meyerhof (1980) as a function of c_1 and the ratio δ / ϕ_1. However, these charts are not presented in non-dimensional form and useful only for the values of sand unit weight and layer thickness that were adopted in the analysis (Burd and Frydman, 1996). Alternatively, useful design charts that cover a broad range of parameters are given by Michalowski and Shi (1995) using limit equilibrium method, which is not experimentally. These solutions, by their very nature, are upper bounds, and they may overestimate the bearing capacity by a significant amount (Burd and Frydman, 1997).

2.6 PREVIOUS ANALYSIS OF BEARING CAPACITY OF FOOTING ON NON-HOMOGENEOUS SOILS

Progress has been made in predicting the bearing capacity on non-homogeneous soils, in particular, the two cases of where the undrained shear strength of the soil varies linearly with depth below the soil surface, i.e. (Poulos et al., 2001):

$$s_u = c_0 + \rho z \tag{2. 24}$$

or below a uniform crust:

$$s_u = c_0 \quad \text{for } z < \frac{c_0}{\rho} \tag{2. 25}$$

$$s_u = \rho z \quad \text{for } z > \frac{c_0}{\rho} \tag{2. 26}$$

where c_0 is the undrained shear strength of the soil at the ground surface, ρ is the strength gradient; and z is the depth below the soil surface. Davis and Booker (1973) used the method of stress characteristics from the theory of plasticity and, assuming the soil obeys the Tresca yield criterion, their plane strain solution was expressed as:

$$q_u = F\left[(2+\pi)c_0 + \frac{\rho B}{4}\right] \tag{2. 27}$$

where F is the bearing capacity factor which is a function of the soil strength non-homogeneity ($\rho B/c_0$) and the roughness of foundation-soil interaction. The bearing capacity factor, F, is reproduced graphically in Figures 2.17 and 2.18 for two different cases of undrained shear strength of the soil, which varies linearly with depth below the soil surface.

NOTE:
This figure is included on page 29 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.17 Bearing Capacity factor, F, for a strip footing on non-homogeneous clay (*After Davis and Booker, 1973*)

As mentioned earlier, soils and rocks are heterogeneous materials created by complex geological processes. The engineering properties of soil vary from point to point, even within the same stratum. Because of the uncertainties associated with their inherent variability, as well as limited information from site investigation, soil and rock properties may be regarded as random variables. Therefore, in the process of engineering analysis and design, the spatial variability of geotechnical properties should be recognized and accounted for (Jaksa, 1995; Paice, et al., 1996; Kaggwa, 2000; Fenton and Griffiths, 2002; 2003; Griffiths and Fenton, 2001).

Figure 2.18 Bearing Capacity factor, *F*, for a strip footing on non-homogeneous clay (*After Davis and Booker, 1973*)

In the past few decades, the effects of heterogeneity of engineering properties on foundation performance have been studied by a number of researchers. For example, with particular relevance to this research, Fenton and Griffiths (2003) examined the bearing capacity of a strip footing located on spatially random c - ϕ soil using elasto-plastic finite element analysis. The soil parameters c and ϕ were simulated as lognormal and correlated random fields. These concepts are treated in the following chapter. In this study, only the Markovian correlation function was used. Fenton and Griffiths (2003) found that, when the soil properties become spatially random, the failure surface passes through the weaker zone or follows the low energy path in the soil beneath the footing, as shown in Figure 2.19, which exhibits bearing failure. Monte Carlo simulation was carried out using different values for various parameters (mean, standard deviation for c and ϕ, correlation distance, and, the correlation between c and ϕ) to study the reliability of Terzaghi's bearing capacity equation. Fenton and Griffiths (2003) concluded that, when c and ϕ were geometrically averaged over some region beneath the footing and were applied to Terzaghi's equation, it gave superior results when estimating the bearing capacity of the strip footing on spatially random soil. The geometric average favours low strength areas, although not as drastically as does a harmonic average, lying in between the arithmetic and harmonic averages. Fenton and Griffiths (2003) also suggested that, in

the absence of site-specific data, a worst-case correlation length, which is approximately equal to the footing width, should be used in order to provide conservative estimates of the probability of bearing failure.

NOTE:
This figure is included on page 31 of the print copy of the thesis held in the University of Adelaide Library.

Figure 2.19 Typical deformed mesh at failure. (*After Fenton and Griffiths, 2003*)

2.7 SUMMARY

Since the introduction of Tezaghi's (1943) bearing capacity equation, extensive research has been carried out in extending the formulation, such that it provides a reasonable estimation of the ultimate bearing capacity of a footing subject to different factors (e.g. footing geometry and load combination). Some research has been carried out in the area of the non-homogeneous and layered soils and rigorous solutions are now available, and the judicious use of their results in practical design is recommended for the following cases:

1. cohesive soil where the undrained shear strength increases linearly with depth (Davis and Booker, 1973);
2. a layer of sand overlying relatively soft clay (Okamura et al., 1998);
3. two layers of cohesive soil (Merifield et al., 1999); and
4. spatially random c - ϕ soil (Fenton and Griffiths, 2003).

This shows that little progress has been made to date in predicting the ultimate bearing capacity of footings on multiple layers of soil. Solving such problem remains beyond the means of simple hand calculation methods. The main reasons for this are

the large number of different cases that may be encountered in practice. This problem requires further investigation, and this is the focus of the work that follows.

CHAPTER 3

NUMERICAL FORMULATION

3.1 INTRODUCTION

In the present study, three different numerical methods have been adopted to determine the ultimate bearing capacity and load-settlement response of foundations. These are upper and lower bound theorems of limit analysis, and more conventional displacement finite element analysis (DFEA). In the first part of this chapter, a general background is provided by briefly discussing these approaches, along with other traditional methods. It is hoped that this will provide an insight into the advantages and disadvantages of current numerical approaches in geomechanics.

The upper and lower bound methods of limit analysis have been used extensively throughout this thesis; therefore, a detailed discussion is appropriate. The use of limit analysis for examining problems in geomechanics is relatively new compared to other methods, thus few geotechnical engineers have a detailed knowledge of this technique. Therefore, some background will be provided to selected aspects of classical plasticity theory, which underpins limit analysis.

In the remaining sections of this chapter, the linear finite element implementations of the upper and lower bound theorems provided by Sloan (1988) and Sloan and Kleeman (1995) are presented. Nonlinear implementation of the lower bound theorem and upper bound theorems by Lyamin and Sloan (2002a, 2002b) is briefly discussed, followed by a discussion on DFEA, each of which will be employed in this study. A brief summary of random field theory is also presented, and in particular, more detailed features of its implementation by means of Local Average Subdivision are also discussed. Finally, the theoretical background information of artificial neural networks (ANNs), which is used in this research, is presented and discussed in detail.

3.2 NUMERICAL METHODS IN GEOMECHANICS

Most geotechnical analysis involves the solution of a boundary value or initial value problem. In order to provide a rigorous solution, four fundamental conditions should be satisfied. These are *equilibrium*, *compatibility*, *constitutive behaviour*, as well as *boundary and initial conditions*.

At present, there are four widely used methods for analysing the bearing capacity of footings. These are the *slip-line, limit equilibrium, limit analysis* and *displacement finite element methods (i.e. DFEMs)*. Traditionally, stability analysis in geotechnical engineering is carried out using the first two of these methods; that is either the slip-line or the limit equilibrium method.

With reference to Table 3.1, all methods, with the exception of the DFEM, fail to satisfy at least one of the four fundamental requirements for the solution of boundary value problems in geotechnical analysis. Only the DFEM provides prediction of the load-deformation response of a deformable solid. Table 3.1 indicates that the DFEM satisfies all the theoretical requirements for a rigorous solution. This method also has the advantage of being able to deal with different and complex variables, such as complicated loadings, excavation and deposition sequences, geometries of arbitrary shape, anisotropy, layered deposits and complex stress-strain relationships. However, great care must be exercised when the DFEM is adopted to predict the true limit load. Sloan and Randolph (1982) showed that the results from the DFEM tend to overestimate the actual limit load. One of the causes of this problem is the additional kinematical constraints imposed on the system by the specified volumetric strains associated with plastic flow (Sloan, 1981; Sloan and Randolph, 1982). Therefore, the DFEM is generally unsuitable for estimating the true limit load.

Alternatively, the limit theorems have the advantage of the ability to bracket the actual collapse load by computing both upper and lower bounds of the solution. These solutions are obtained using numerical techniques, such as those developed by Sloan (1988) and Sloan and Kleeman (1995), which are based on the limit theorems of classical plasticity and finite elements.

Deriving statically-admissible stress fields or kinematically-admissible velocity fields by hand is especially difficult for problems involving complicated geometries, heterogeneous materials, or complex loading. Hence, computer-based numerical methods are usually required. The most common numerical implementation of the lower and upper bound theorem is based on a finite element discretisation of the continuum and results in an optimisation problem with large, sparse constraint matrices. The following section discusses these methods in some detail.

Table 3.1 Comparison of existing methods of analysis. (*After Merifield, 2002*)

NOTE:
This table is included on page 36 of the print copy of the thesis held in the University of Adelaide Library.

3.3 THEORY OF LIMIT ANALYSIS

Various elementary ideas of plastic deformation and failure, and the reduction of buckling loads gradually emerged throughout the nineteenth century. During that time, the general concepts of plasticity were established by Shanley (1947), Hill (1950), Drucker (1950), Drucker et al. (1951), Prager and Hodge (1951), Symonds and Neal (1951), Koiter (1953) and others. One of the most significant developments was made when Drucker et al. (1952) established the upper and lower bound limit analyses based on the plastic limit theorems. This was further refined by Chen (1975). Since then, limit analysis has become increasingly popular for predicting continuum collapse loads.

It is extremely difficult, if not impossible, to derive statistically-admissible stress fields manually for complicated problems. Consequently, efforts have been made in solving limit analysis problems by using the finite element method, which represents an attempt to obtain a lower or upper bound solution by numerical methods on a theoretically rigorous foundation of plasticity. It was first proposed by Lysmer (1970) and followed by others, e.g. Anderheggen and Knopfel (1972); Pastor (1978); Bottero et al. (1980). Sloan and his co-workers (Sloan, 1988, 1989; Sloan and Kleeman, 1995) have made significant progress in developing the methods using finite elements

and linear programming (LP) for computing rigorous lower and upper bound solutions for 2D stability problems, mainly in relation to bearing capacity. Lyamin and Sloan (2002a, 2002b) further developed lower and upper bound limit analysis using finite elements and nonlinear programming (NLP) for 3D stability problems. Their numerical implementations are based on the finite element discretization of the rigid plastic continuum and results in a standard linear or nonlinear optimization problem with a highly sparse set of constraints. The numerical implementation is also based on the yield criterion, flow rule, boundary conditions, and the energy–work balance equation.

The limit theorems provide a simple and useful way of analysing the stability of structures, and they have been applied to the problem of the bearing capacity of a footing on layered soil. For example, as mentioned previously, Merifield et al. (1999) conducted a parametric study involving the bearing capacity of the footing on a two-layered clay using finite element limit analysis and, more recently, Shiau et al. (2003) applied limit analysis to determine the upper and lower bound solution of the bearing capacity of a footing founded on a dense sand overlying a soft clay layer.

Limit analysis methods assume a *perfectly plastic* model with an *associated flow rule*, which implies that the plastic strain rates are normal to the yield surface. The validity of these methods, or any other analysis method, depends on the underlying assumptions made. These assumptions are reviewed in detail in the following section.

3.3.1 Perfectly Plastic Material

In the limit theorems, an elastic-perfectly-plastic model is adopted to describe the load-deformation behaviour of soil. The strain-softening features, which are typically exhibited in overconsolidated clays and dense sands are simplified under such an assumption. When plastic deformation occurs, an elastic-perfectly-plastic material exhibits the property of continuing plastic deformation or flow at constant stress. In limit analysis, the deformation at collapse occurs in the *rigid perfectly plastic* stage, and therefore the elastic properties play no part in the analysis process.

3.3.2 Yield Criterion

For a perfectly plastic material, the yield function (f) depends on the material strength and invariant combination of stress components, σ_{ij}. The definition of the yield function implies that $f(\sigma_{ij})$ is 0 at the yield surface, when plastic flow occurs, and $f(\sigma_{ij}) < 0$ when it is within the yield surface, when elastic behaviour is observed. Positive values of $f(\sigma_{ij})$ imply stresses lying outside the yield surface which are illegal or inadmissible.

3.3.3 Stability Postulate

Drucker et al. (1952) introduced the idea of a *stable plastic material,* as follows:

1. A material can be classified as *stable* if the stress is a unique function of strain, and vice versa.
2. A material can be classified as *unstable* if the stress is not a unique function of strain, and vice versa.

According to the definitions above, Drucker et al. (1952) expressed the concept of the stability postulate in vector form:

$$(\sigma - \sigma^0)^T \dot{\varepsilon}^p \geq 0 \tag{3.1}$$

where σ^0 is the initial stress vector, σ is the final stress vector and $\dot{\varepsilon}^p$ is the vector of the plastic strain rate. The parameters σ^0, σ and $\dot{\varepsilon}^p$ are defined as:

$$\sigma^T = \left\{ \sigma_x, \sigma_y, \sigma_z, \tau_{xy}, \tau_{yz}, \tau_{zx} \right\} \tag{3.2}$$

$$\sigma^{0\,T} = \left\{ \sigma_x^0, \sigma_y^0, \sigma_z^0, \tau_{xy}^0, \tau_{yz}^0, \tau_{zx}^0 \right\} \tag{3.3}$$

$$\dot{\varepsilon}^{p\,T} = \left\{ \dot{\varepsilon}_x^p, \dot{\varepsilon}_y^p, \dot{\varepsilon}_z^p, \dot{\gamma}_{xy}^p, \dot{\gamma}_{yz}^p, \dot{\gamma}_{zx}^p \right\} \tag{3.4}$$

3.3.4 Flow Rule

When kinematics of plastic flow are considered, the process of plastic flow is described in terms of the strain rate, $\dot{\varepsilon}_{ij}$. For a yielding perfectly-plastic material, the total strain rate, $\dot{\varepsilon}_{ij}$, contains both elastic and plastic components:

$$\dot{\varepsilon}_{ij} = \dot{\varepsilon}_{ij}^{e} + \dot{\varepsilon}_{ij}^{p} \tag{3.5}$$

where $\dot{\varepsilon}_{ij}^{e}$ is the elastic strain rate. During plastic deformation, the material may flow in an *associated* manner; that is, the plastic strain rates are normal to the failure surface. Therefore, the axes of the principal plastic strain rate correspond to the axes of principal stresses.

A *stable* material that satisfies the Drucker's postulate must have the following properties (Drucker, 1950):

1. The yield surface $f(\sigma_{ij})$ must be convex.
2. The plastic strain rate must be normal to the yield surface $\dot{\varepsilon}_{ij}^{p} = \dot{\lambda}\dfrac{\partial f}{\partial \sigma_{ij}}$, where $\dot{\lambda} \geq 0$ is known as the plastic multiplier rate.

It is proved that convexity and normality are sufficient to satisfy Drucker's postulate.

3.3.5 Small Deformations and the Principle of Virtual Work

In the theorem of limit analysis, the body does not undergo large changes in its geometry at collapse, so that the equations of virtual work are applicable. These equations consist of two sets of variables, which are those defining an equilibrium stress field and those defining a compatible deformation field, and can be expressed as:

$$\int_A T_i \dot{u}_i \, \mathrm{d}A + \int_V F_i \dot{u}_i \, \mathrm{d}V = \int_V \sigma_{ij} \dot{\varepsilon}_{ij} \, \mathrm{d}V \tag{3.6}$$

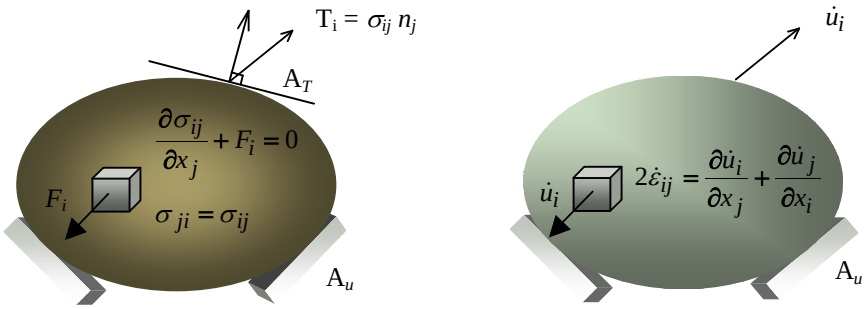
In limit analysis, the equilibrium stress and compatible deformation fields are not required to correspond to the actual state, nor are related to one another. In Equation 3.6, the integration is over the surface area, *A*, and volume, *V*, of the body. The tensor σ_{ij} is any set of stresses real, or otherwise, that is in equilibrium with the body force, F_i, and the external surface tractions, T_i. With reference to Figure 3.1(a), an equilibrium stress field must satisfy the following equations:

$$\text{At surface points:} \quad T_i = \sigma_{ij} n_j ; \tag{3.7}$$

$$\text{At interior points:} \quad \frac{\partial \sigma_{ij}}{\partial x_j} + F_j = 0 \text{ ; and} \tag{3.8}$$

$$\sigma_{ji} = \sigma_{ij} \tag{3.9}$$

where n_i is the outward normal to a surface element at any point.



(a) (b)

Figure 3.1 Stress and deformation fields in the equation of virtual work: (a) Equilibrium stress field; and (b) Compatible deformation field

In Equation 3.6, the strain rate, $\dot{\varepsilon}$, denotes any set of strains, compatible with real or imagined (virtual) displacement rate, $\dot{u}$, which arises from the application of F_i and T_i. Referring to Figure 3.1, a compatible deformation field must satisfy the compatibility relation (Hill, 1950):

$$2\dot{\varepsilon}_{ij} = \frac{\partial \dot{u}_i}{\partial x_j} + \frac{\partial \dot{u}_j}{\partial x_i} \tag{3.10}$$

where x_i represents the position vectors referred to Cartesian coordinates fixed in the element.

3.3.6 The Limit Theorems

Following the above brief review of the assumptions implicit in the limit analysis theorems, a summary can be made as follows:

1. The material is perfectly plastic, so that no work-hardening nor work-softening is permitted. This implies that the stress point cannot move outside the yield surface, which remains fixed in stress space, and that the vector $\dot{\sigma}_{ij}$ must be tangential to this surface whenever plastic straining occurs.
2. The plastic strain rates are normal to the yield surface and are given by an associated flow rule of the form $\dot{\varepsilon}_{ij}^{p} = \lambda \delta f / \delta \sigma_{ij}$. From the assumption of perfect plasticity and normality, it follows that $\dot{\sigma}_{ij}\ \dot{\varepsilon}_{ij}^{p} = 0$.
3. The yield surface is convex.
4. The body does not undergo large changes in its geometry at collapse, so that the equations of virtual work are applicable.

As mentioned before, all deformations at collapse are purely plastic, and this feature implies that the elastic properties play no part in collapse. The limit load is defined as the plastic collapse load of a body having the idealised properties given in (1) to (3), above.

The lower bound limit theorem may be summarised as follows (Drucker et al., 1952):

If a stress distribution σ_{ij}^{s} can be found which satisfies equilibrium, balances the applied traction, T_i, on the boundary, A_T, and does not violate the yield criterion, so that $f(\sigma_{ij}^{s}) \leq 0$, then the tractions and body force, F_i, will be less than or equal to the actual tractions and body forces that cause collapse.

To prove the validity of this theorem, let us suppose that there are two stress fields: the first is the actual collapse state of a body, which is expressed in terms of traction, T_i^c, and body force, F_i^c, (Equation 3.11); and the second is another stress field which supports the tractions, T_i^s, and the body forces, F_i^s, (Equation 3.12). We also assume proportional loading with scalar multipliers λ_T^s and λ_F^s, which slowly increase from zero, so that $T_i^s = \lambda_T^s T_i^c$ and $F_i^s = \lambda_F^s F_i^c$. Since all deformation at collapse is plastic, therefore $\dot{\varepsilon}_{ij}^c = \dot{\varepsilon}_{ij}^{pc}$ and $\dot{u}_{ij}^c = \dot{u}_{ij}^{pc}$, and the virtual work equations for this system are:

$$\int_{A_T} T_i^c \dot{u}_i^{pc} \mathrm{d}A + \int_V F_i^c \dot{u}_i^{pc} \mathrm{d}V = \int_V \sigma_{ij}^c \dot{\varepsilon}_{ij}^{pc} \mathrm{d}V \tag{3.11}$$

$$\int_{A_T} (T_i^s = \lambda_T^s T_i^c) \dot{u}_i^{pc} \mathrm{d}A + \int_V (F_i^s = \lambda_F^s F_i^c) \dot{u}_i^{pc} \mathrm{d}V = \int_V \sigma_{ij}^s \dot{\varepsilon}_{ij}^{pc} \mathrm{d}V \tag{3.12}$$

Subtracting Equation 3.12 from Equation 3.11 yields:

$$(1-\lambda_T^s) \int_{A_T} T_i^c \dot{u}_i^{pc} \mathrm{d}A + (1-\lambda_F^s) \int_V F_i^c \dot{u}_i^{pc} \mathrm{d}V = \int_V (\sigma_{ij}^c - \sigma_{ij}^s) \dot{\varepsilon}_{ij}^{pc} \mathrm{d}V \tag{3.13}$$

The conditions of convexity and normality imply that $(\sigma_{ij}^c - \sigma_{ij}^s)\, \dot{\varepsilon}_{ij}^{pc} \geq 0$ for all points in the body. Thus Equation 3.13 becomes:

$$(1-\lambda_T^s) \int_{A_T} T_i^c \dot{u}_i^{pc} \mathrm{d}A + (1-\lambda_F^s) \int_V F_i^c \dot{u}_i^{pc} \mathrm{d}V \geq 0 \tag{3.14}$$

For cases where the body forces are fixed, the second integral in Equation 3.14 vanishes. Thus, it follows that $\lambda_T^s \leq 1$. Conversely, if all surface tractions are fixed and therefore the first integral vanishes, and Equation 3.14 implies that $\lambda_T^c \leq 1$. This result proves that the load supported by σ_{ij}^s is never greater than or equal to the true collapse load and hence establishes the lower bound theorem. For the principles of

limit analysis, any stress field is said to be *statically admissible* if it is in equilibrium with the surface tractions, satisfies the stress boundary condition and nowhere violates the yield condition.

The upper bound limit theorem may be stated as follows (Drucker et al., 1952):

If a compatible plastic deformation field $(\dot{\varepsilon}_{ij}^{pk}, \dot{u}_i^{pk})$ *can be found which satisfies the velocity boundary condition,* $\dot{u}_i^{pk} = 0$*, on the boundary and the normality condition* $\dot{\varepsilon}_{ij}^{pk} = \lambda \delta f / \delta \sigma_{ij}$ *, then the tractions and body force determined by equating the rate of work of the external forces* (Equation 3.15) *to the internal dissipation rate* (Equation 3.16) *will be greater than or equal to the actual tractions and body forces that cause collapse.*

$$\int_{A_T} T_i \, \dot{u}_i^{pk} \mathrm{d}A + \int_V F \, \dot{u}_i^{pk} \mathrm{d}V \tag{3.15}$$

$$\int_V D \cdot \dot{\varepsilon}_{ij}^{pk} \mathrm{d}V = \int_V \sigma_{ij}^{pk} \, \dot{\varepsilon}_{ij}^{pk} \mathrm{d}V \tag{3.16}$$

To prove this theorem, as for the lower bound theorem, let us suppose that the actual collapse state of a body is described by the tractions T_i^c, body forces F_i^c and stresses σ_{ij}^c. The virtual work (Equation 3.6) can be written as:

$$\int_{A_T} T_i^c \, \dot{u}_i^{pk} \mathrm{d}A + \int_V F_i^c \, \dot{u}_i^{pk} \mathrm{d}V = \int_V \sigma_{ij}^c \, \dot{\varepsilon}_{ij}^{pk} \mathrm{d}V \tag{3.17}$$

We further suppose that T_i^k and F_i^k are estimates of the true collapse tractions and body forces, which are obtained by substituting the strains $\dot{\varepsilon}_{ij}^{pk}$, the stresses σ_{ij}^{pk} and the displacement rates $\dot{u}_{ij}^{pk}$ from a compatible plastic deformation field into

Equations 3.15 and 3.16. Again, we assume proportional loading with $T_i^k = \lambda_T^k T_i^c$ and $F_i^k = \lambda_F^k F_i^c$. Substituting these quantities into virtual work Equation 3.6 gives:

$$\int_{A_T} (T_i^k = \lambda_T^k T_i^c)\dot{u}_i^{pk} \mathrm{d}A + \int_V (F_i^k = \lambda_F^k F_i^c)\dot{u}_i^{pk} \mathrm{d}V = \int_V \sigma_{ij}^{pk}\, \dot{\varepsilon}_{ij}^{pk} \mathrm{d}V \qquad (3.18)$$

Subtracting Equation 3.18 from Equation 3.17 gives:

$$(\lambda_T^k - 1)\int_{A_T} T_i^c\, \dot{u}_i^{pk} \mathrm{d}A + (\lambda_F^k - 1)\int_V F_i^c\, \dot{u}_i^{pk} \mathrm{d}V \geq \int_V (\sigma_{ij}^{pk} - \sigma_{ij}^c)\, \dot{\varepsilon}_{ij}^{pk} \mathrm{d}V \qquad (3.19)$$

The conditions of convexity and normality imply that $(\sigma_{ij}^{pk} - \sigma_{ij}^c) \cdot \dot{\varepsilon}_{ij}^{pk} \geq 0$ for all points in the body. Substituting this inequality into Equation 3.19 yields the upper bound relation:

$$(\lambda_T^k - 1)\int_{A_T} T_i^c\, \dot{u}_i^{pk} \mathrm{d}A + (\lambda_F^k - 1)\int_V F_i^c\, \dot{u}_i^{pk} \mathrm{d}V \geq 0 \qquad (3.20)$$

For cases where the body forces, F_i^c, are fixed, the second integral in Equation 3.20 disappears. Thus, it follows that $\lambda_T^k \geq 1$. Conversely, if all surface tractions, T_i^c, are fixed, the first integral vanishes, and Equation 3.20 implies that $\lambda_F^k \geq 1$. This result proves that the load supported by σ_{ij}^{pk} is never less than or equal to the true collapse load and hence establishes the upper bound theorem. Any compatible deformation field is said to be kinematically admissible, if one satisfies the velocity boundary conditions and the flow rule. In contrast to the lower bound theorem, the upper bound theorem is concerned only with the kinematic variables and their associated energy dissipation.

The lower and upper bound theorems can also be applied to discontinuous fields of stress and velocity, again, because they yield reliable estimates of the true collapse load. Surfaces of stress discontinuity are admissible if the equilibrium equations are satisfied at all points on these surfaces. Also, surfaces of velocity discontinuity can

also be admitted, provided that the energy dissipation is properly computed. Velocity discontinuities are essential in mechanisms comprised of rigid blocks, where energy is dissipated solely by sliding between adjacent rigid body blocks. Conceptually, velocity discontinuities may be viewed as the limiting case of a continuous velocity field, in which one or more velocity components change very rapidly across a narrow transition layer, which is replaced by a discontinuity of zero thickness. The discontinuous velocity fields are often observed in actual collapse mechanisms, in contrast to stress discontinuities, which are useful, but rarely resemble the actual stress state.

The lower bound technique is inherently more attractive than the upper bound technique, since it automatically provides a conservative estimate of the true collapse load. However, in practice, it is appropriate to use the upper and lower bound methods in parallel to bracket the true limit load from above and below, as the difference between the bounds provides an in-built error indicator, which should be as small as possible (Hjiaj et al. 2003).

The following sections explain in greater detail the derivation of the lower bound, followed by the upper bound limit analysis formulation.

3.4 LOWER BOUND LIMIT ANALYSIS FORMULATION

The sign conventions adopted for the stresses in the lower bound limit formulation are shown in Figure 3.2, with tension taken as positive. Three types of elements are employed, as illustrated in Figure 3.3, and each permits the stresses to vary linearly according to (Sloan, 1988):

$$\sigma_x = \sum_{i=1}^{i=3} N_i \sigma_{xi} \tag{3.21}$$

$$\sigma_y = \sum_{i=1}^{i=3} N_i \sigma_{yi} \tag{3.22}$$

$$\tau_{xy} = \sum_{i=1}^{i=3} N_i \tau_{xyi} \tag{3.23}$$

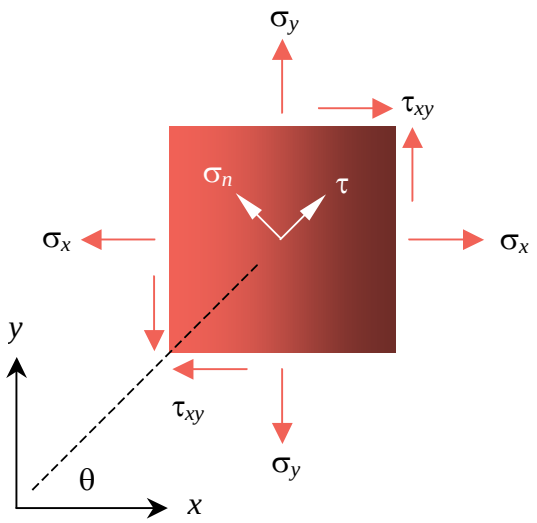


Figure 3.2 Stress sign convention. (*After Sloan, 1988*)

where N_i is linear shape function and can be expressed in terms of element nodal coordinates (x_i , y_i) as follows:

$$N_1 = [(x_2y_3 - x_3y_2) + y_{23}x + x_{32}y] / 2A \tag{3.24}$$

$$N_2 = [(x_3y_1 - x_1y_3) + y_{31}x + x_{13}y] / 2A \tag{3.25}$$

$$N_3 = [(x_1y_2 - x_2y_1) + y_{12}x + x_{21}y] / 2A \tag{3.26}$$

where:

$$x_{32} = x_3 - x_2 \qquad y_{32} = y_3 - y_2$$

$$x_{13} = x_1 - x_3 \qquad y_{13} = y_1 - y_3$$

$$x_{21} = x_2 - x_1 \qquad y_{21} = y_2 - y_1$$

and: $$2A = \left| x_{13}y_{23} - x_{32}y_{31} \right|$$

and A is the triangular area. Note that the rectangular and triangular extension elements, which enable a statically admissible stress field to be obtained for a semi-infinite domain, are based on the same linear expansion as the 3-noded triangle. Unlike the displacement finite element methods (DFEM), which is discussed later,

each node is unique to a single element in the lower bound mesh, although several nodes may share the same coordinates. Statically admissible stress discontinuities are applied at all edges that are shared by adjacent elements.

A rigorous lower bound of the exact collapse load is ensured by insisting that the stresses obey equilibrium, and satisfy both the stress boundary conditions and yield criterion. Each of these requirements imposes a separate set of constraints on the nodal stresses.

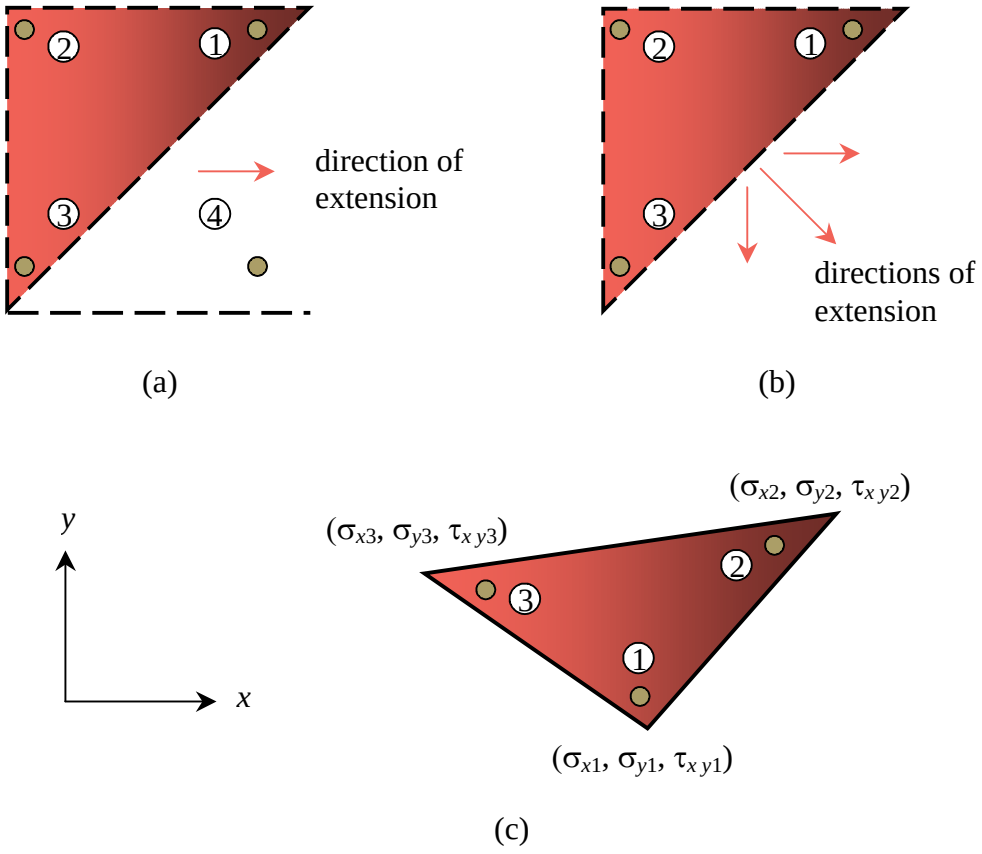


Figure 3.3 Element types for lower bound analysis. (*After Sloan, 1988*) (a) 4-noded rectangular extension element; (b) 3-noded triangular extension element; and (c) 3-noded triangular element

3.4.1 Constraints from Equilibrium Conditions

With reference to Figure 3.2, the equilibrium conditions for plane strain are:

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} = 0 \tag{3.27}$$

$$\frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{xy}}{\partial x} = \gamma \tag{3.28}$$

Substituting Equations 3.21 to 3.23 into the equilibrium equations (Equation 3.27 and and 3.28), the nodal stresses for each element subject to two equilibrium constraints can be depicted as follows:

$$\begin{bmatrix} y_{23} & 0 & x_{32} & y_{31} & 0 & x_{13} & y_{12} & 0 & x_{21} \\ 0 & x_{32} & y_{23} & 0 & x_{13} & y_{31} & 0 & x_{21} & y_{12} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{x1} \\ \sigma_{y1} \\ \tau_{xy1} \\ \sigma_{x2} \\ \sigma_{y2} \\ \tau_{xy2} \\ \sigma_{x3} \\ \sigma_{y3} \\ \tau_{xy3} \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma \end{bmatrix} \tag{3.29}$$

which can be expressed in the form of:

$$\mathbf{a_1\, x = b_1} \tag{3.30}$$

For the rectangular extension element case, three addition equalities are necessary to extend the linear stress distribution to the fourth node. These equalities are:

$$\sigma_{x4} = \sigma_{x1} + \sigma_{x3} - \sigma_{x2} \tag{3.31}$$

$$\sigma_{y4} = \sigma_{y1} + \sigma_{y3} - \sigma_{y2} \tag{3.32}$$

$$\tau_{xy4} = \tau_{xy1} + \tau_{xy3} - \tau_{xy2} \tag{3.33}$$

and may be written as:

$$\begin{bmatrix} T_1 & -T_1 & T_1 & -T_1 \end{bmatrix} \cdot \begin{bmatrix} \sigma_{x1} \\ \sigma_{y1} \\ \tau_{xy1} \\ \vdots \\ \sigma_{x4} \\ \sigma_{y4} \\ \tau_{xy4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.34}$$

where:

$$T_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which can be expressed in the form of:

$$\mathbf{a_2\, x = 0}$$

The fourth node of the rectangular extension element is essentially a dummy node but is necessary to permit semi-infinite stress discontinuities between adjacent extension elements.

3.4.2 Constraints from Stress Discontinuity

A stress discontinuity is statically admissible if the shear and normal stresses acting on the discontinuity plane are continuous (only the tangential stress may jump across elements). With reference to Figure 3.2, the normal and shear stresses acting on a plane at an angle θ to the *x*-axis are given by the stress transformation equations:

$$\sigma_n = \sigma_x \sin^2\theta + \sigma_y \cos^2\theta - \tau_{xy} \sin 2\theta \tag{3.35}$$

$$\tau = \frac{1}{2} (\sigma_y - \sigma_x) \sin 2\theta + \tau_{xy} \cos 2\theta \tag{3.36}$$

A typical stress discontinuity between adjacent triangles is shown in Figure 3.4. It is defined by the nodal pairs (1, 2) and (3, 4), where the nodes in each pair have identical coordinates. Since the stresses in the model are assumed to vary linearly, the equilibrium condition is met by equalising shear and normal stresses of all pairs of nodes on opposite sides of the discontinuity. With reference to Figure 3.4, these constraints may be written as:

$$\sigma_{n1} = \sigma_{n2}; \quad \sigma_{n3} = \sigma_{n4}; \quad \tau_1 = \tau_2 \quad \text{and} \quad \tau_3 = \tau_4 \tag{3.37}$$

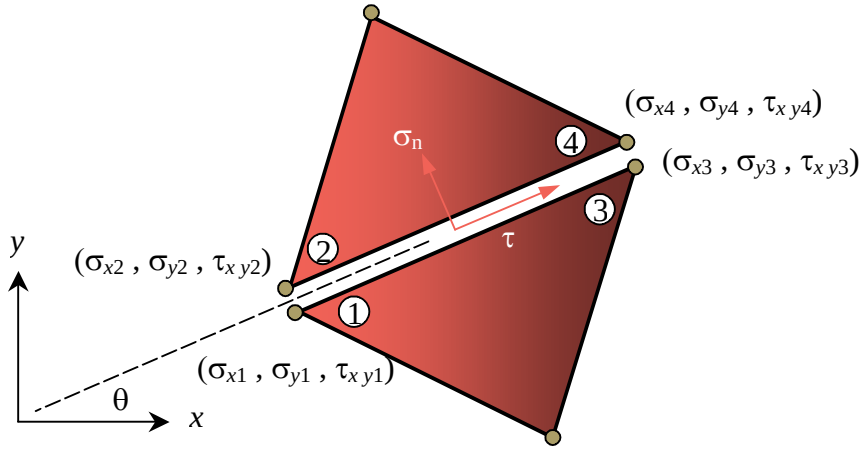


Figure 3.4 Stress discontinuity. (*After Sloan, 1988*)

Substituting Equation 3.35 and 3.36 into Equation 3.37 above, the discontinuity equilibrium condition becomes:

$$\begin{bmatrix} T_2 & -T_2 & 0 & 0 \\ 0 & 0 & T_2 & -T_2 \end{bmatrix} \cdot \begin{bmatrix} \sigma_{x1} \\ \sigma_{y1} \\ \tau_{xy1} \\ \vdots \\ \sigma_{x4} \\ \sigma_{y4} \\ \tau_{xy4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.38}$$

where

$$T_2 = \begin{bmatrix} \sin^2\theta & \cos^2\theta & -\sin 2\theta \\ -\frac{1}{2}\sin 2\theta & \frac{1}{2}\sin 2\theta & \cos 2\theta \end{bmatrix} \tag{3.39}$$

which can be written in form of:

$$\mathbf{a}_3\,\mathbf{x} = \mathbf{0} \tag{3.40}$$

3.4.3 Constraints from Stress Boundary Conditions

To implement the prescribed boundary conditions, it is necessary to impose additional equality constraints on the nodal stresses. If the normal and shear stresses at the ends of a boundary segment are specified to be (q_1, t_1) and (q_2, t_2), as shown in Figure 3.5, then it is sufficient to impose the following conditions:

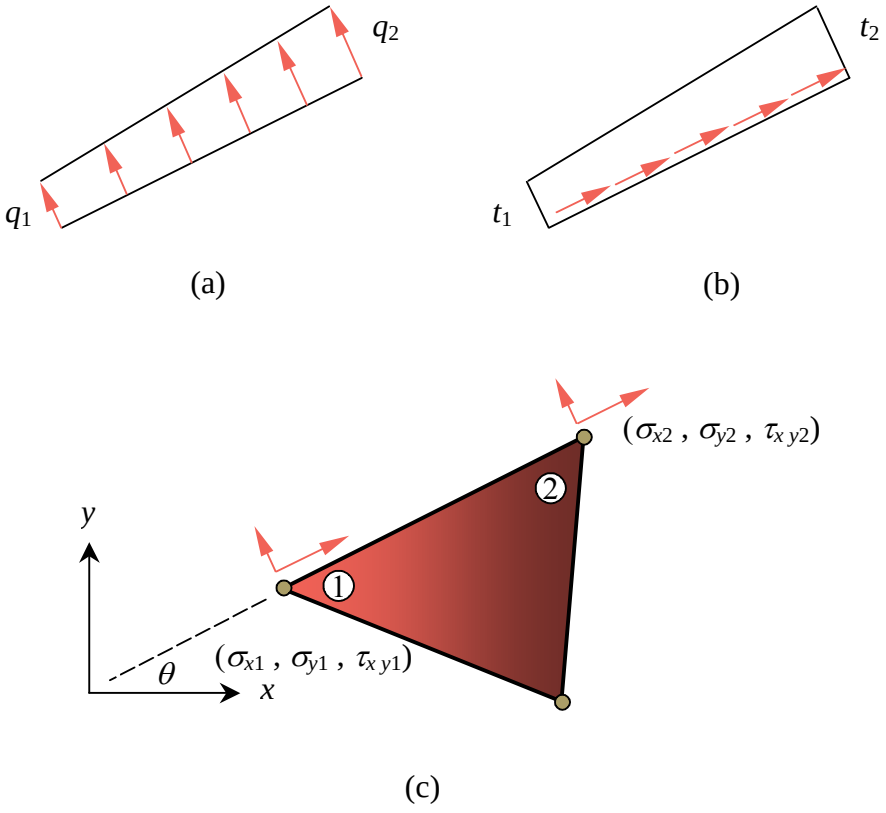


Figure 3.5 Stress boundary conditions: (a) prescribed normal stresses, (b) prescribed shear stresses, and (c) 3-noded triangular element with additional equality constraints.

$$\sigma_{n1} = q_1 ; \quad \sigma_{n2} = q_2 ; \quad \tau_1 = t_1 \quad \text{and} \quad \tau_2 = t_2 \tag{3.41}$$

since the stresses are only permitted to vary linearly along an element edge.

Substituting the stress transformation Equations 3.35 and 3.36 into Equation 3.41 above, leads to four equalities of the general form:

$$\begin{bmatrix} \mathrm{T}_2 & 0 \\ 0 & \mathrm{T}_2 \end{bmatrix} . \begin{bmatrix} \sigma_{x1} \\ \sigma_{y1} \\ \tau_{xy1} \\ \sigma_{x2} \\ \sigma_{y2} \\ \tau_{xy2} \end{bmatrix} = \begin{bmatrix} q_1 \\ t_1 \\ q_2 \\ t_2 \end{bmatrix} \tag{3.42}$$

which can be written as:

$$\mathbf{a_4\, x = b_4} \tag{3.43}$$

Note that Equation 3.42 can also be applied to an extension element to ensure that the stress boundary conditions are satisfied everywhere along a semi-infinite edge.

In cases involving a uniform (but unknown) applied surface traction, for example in the analysis of flexible foundations, it is necessary to place additional constraints on the unknown stresses, which are of the form:

$$\sigma_{n1} = \sigma_{n2} \quad \text{and} \quad \tau_1 = \tau_2 \tag{3.44}$$

Substituting the stress transformation Equations 3.35 and 3.36 into Equation 3.44 above leads to two equalities of the form:

$$\begin{bmatrix} \mathrm{T}_2 & -\mathrm{T}_2 \end{bmatrix} \cdot \begin{bmatrix} \sigma_{x1} \\ \sigma_{y1} \\ \tau_{xy1} \\ \sigma_{x2} \\ \sigma_{y2} \\ \tau_{xy2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.45}$$

which can be written as:

$$\mathbf{a_5\ x = 0} \tag{3.46}$$

3.4.4 Constraints from Yield Conditions

For plane strain conditions, the Mohr-Coulomb yield criterion is presented as:

$$F = (\sigma_x - \sigma_y)^2 + (2\tau_{xy})^2 - (2c\cos\phi - (\sigma_x + \sigma_y)\sin\phi)^2 = 0 \tag{3.47}$$

where c is the soil cohesion and ϕ is the soil friction angle. For an internal linearisation with p number of planes, the k^{th} plane of the Mohr-Coulomb criterion is as follows:

$$F_k = A_k\sigma_x + B_k\sigma_y + C_k\tau_{xy} - 2c\cos\phi\cos(\pi/p) = 0 \tag{3.48}$$

where:

$$A_k = \cos(2\pi\kappa/p) + \sin\phi\cos(\pi/p)$$

$$B_k = \sin\phi\cos(\pi/p) - \cos(2\pi\kappa/p) \tag{3.49}$$

$$C_k = 2\sin(2\pi\kappa/p)$$

and $k = 1, 2, \ldots, p$. It is suggested that in order to model the Mohr-Coulomb yield function with sufficient accuracy, the minimum requirement of p is 12, and higher values may be needed for soils with high friction angles (Sloan, 1988).

Each side of the polygonal yield surface given by Equation 3.48 is expressed as a linear function of the unknown stresses. This equation needs to be enforced at each node of each triangular element to ensure that the stresses satisfy the yield condition everywhere. Let F_{ki} denotes the value of Mohr-Coulomb yield function of the k^{th} side of the yield function at node i, then linearised yield constraint may be written as $F_{ki} \leq 0$, where $i = 1, 2, 3$ and $k = 1, 2, \ldots, p$. These constraints give rise to a set of p number of inequalities for each node, i, of the form:

$$\begin{bmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ \vdots & \vdots & \vdots \\ A_k & B_k & C_k \\ \vdots & \vdots & \vdots \\ A_p & B_p & C_p \end{bmatrix} \cdot \begin{bmatrix} \sigma_{xi} \\ \sigma_{yi} \\ \tau_{xyi} \end{bmatrix} \leq \begin{bmatrix} 2c_i \cos\phi \cos(\pi / p) \\ 2c_i \cos\phi \cos(\pi / p) \\ \vdots \\ 2c_i \cos\phi \cos(\pi / p) \\ \vdots \\ 2c_i \cos\phi \cos(\pi / p) \end{bmatrix} \tag{3.50}$$

where c_i denotes the cohesion at node i and can be expressed in form of:

$$\mathbf{a_6\, x \leq b_6} \tag{3.51}$$

For the four-noded rectangular extension element, the constraints can be expressed as:

$$\begin{bmatrix} A_1 & B_1 & C_1 & -A_1 & -B_1 & -C_1 \\ A_2 & B_2 & C_2 & -A_2 & -B_2 & -C_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_k & B_k & C_k & -A_k & -B_k & -C_k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_p & B_p & C_p & -A_p & -B_p & -C_p \end{bmatrix} \cdot \begin{bmatrix} \sigma_{xi} \\ \sigma_{yi} \\ \tau_{xyi} \end{bmatrix} \leq \begin{bmatrix} 2(c_1 - c_2) \cos\phi \cos(\pi / p) \\ 2(c_1 - c_2) \cos\phi \cos(\pi / p) \\ \vdots \\ 2(c_1 - c_2) \cos\phi \cos(\pi / p) \\ \vdots \\ 2(c_1 - c_2) \cos\phi \cos(\pi / p) \end{bmatrix} \tag{3.52}$$

and:

$$\mathbf{a_7\, x \leq b_7} \tag{3.53}$$

3.4.5 Formation of Lower Bound Objective Function

For two-dimensional plane strain geotechnical stability problems, a statically admissible stress field is sought, which maximises the integral of the normal stress over some part of the boundary. These integrals may be represented as:

$$Q_u = h\int \sigma_n ds \tag{3.54}$$

where Q_u represents the collapse load and h is the out-of-plane thickness (Sloan, 1988). The integration can be performed analytically and after substitution of the stress transformation equations, the collapse loads may be written as:

$$\boldsymbol{Q_u = c^T x} \tag{3.55}$$

where $\mathbf{c}^T$ is known as the objective function, since it defines the quantity which is to be optimised. Once the elemental constraint matrices and objective function coefficients have been found, the various terms may be assembled to furnish the lower bound programming problem, which is written as:

$$\begin{aligned} &\textit{Maximize:} && \mathbf{c_1}^T\mathbf{x} \\ &\textit{Subject to:} && \mathbf{a_1 x = b_1} \\ & && \mathbf{a_2 x = 0} \\ & && \mathbf{a_3 x = 0} \\ & && \mathbf{a_4 x = b_4} \\ & && \mathbf{a_5 x = 0} \\ & && \mathbf{a_6 x \le b_6} \\ & && \mathbf{a_7 x \le b_7} \end{aligned} \tag{3.56}$$

3.5 UPPER BOUND LIMIT ANALYSIS FORMULATION

A three-noded triangular element used in the upper bound formulation is shown in Figure 3.6. There are two velocity components at each node and each element has *p*

plastic multiplier rates $\dot{\lambda}_i$, where p is the number of planes in the linearised yield criterion.

With reference to Figure 3.6, the velocities within the triangle are assumed to vary linearly according to:

$$u = \sum_{i=1}^{i=3} N_i u_i \tag{3.57}$$

$$v = \sum_{i=1}^{i=3} N_i v_i \tag{3.58}$$

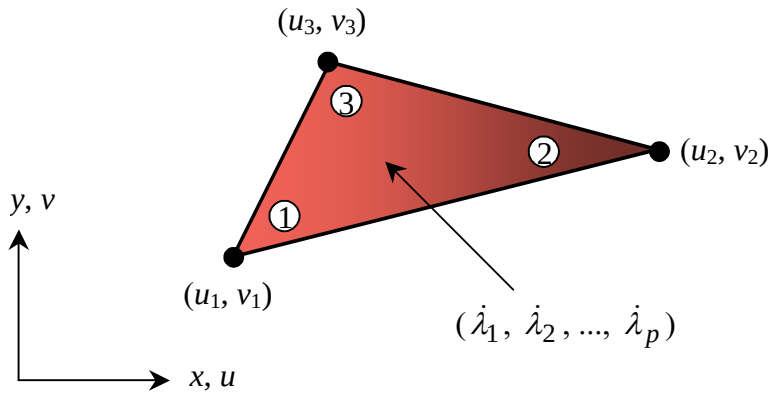


Figure 3.6 Three-noded triangular element. (*After Sloan and Kleeman, 1995*)

where u and v are nodal velocity in the x-, y- directions, respectively, N_i are linear shape functions defined in Equations 3.24 to 3.26, and may be expressed in terms of (x_i, y_i) nodal coordinates.

3.5.1 Constraints from Plastic Flow in Continuum

To be kinematically admissible, the velocity field must satisfy the set of constraints imposed by an associated flow rule. For plane strain deformation of rigid plastic soil, the associatied flow rule is given in the form:

$$\dot{\varepsilon}_x = \frac{\partial u}{\partial x} = \dot{\lambda}\frac{\partial F}{\partial \sigma_x} \tag{3.59}$$

$$\dot{\varepsilon}_y = \frac{\partial v}{\partial y} = \dot{\lambda}\frac{\partial F}{\partial \sigma_y} \tag{3.60}$$

$$\dot{\gamma}_{xy} = \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) = \dot{\lambda}\frac{\partial F}{\partial \tau_{xy}} \tag{3.61}$$

where tensile strains are taken as positive and these equations contain stress terms, which must be removed to provide a linear relationship between the unknown velocities and plastic multiplier rates. The removal of the stress terms can be achieved by employing a linear approximation to the yield surface. For an external linearisation with p number of planes, the k^{th} plane of the Mohr-Coulomb criterion is as follows:

$$F_k = A_k\sigma_x + B_k\sigma_y + C_k\tau_{xy} - 2c\cos\phi = 0 \tag{3.62}$$

where:

$$A_k = \cos(2\pi\kappa/p) + \sin\phi$$

$$B_k = \sin\phi - \cos(2\pi\kappa/p) \tag{3.63}$$

$$C_k = 2\sin(2\pi\kappa/p)$$

Substituting Equation 3.62 into Equations 3.59, 3.60 and 3.61, the plastic strain rates may be shown as:

$$\dot{\varepsilon}_x = \frac{\partial u}{\partial x} = \sum_{k=1}^{k=p}\dot{\lambda}_k\frac{\partial F_k}{\partial \sigma_x} = \sum_{k=1}^{k=p}\dot{\lambda}_k A_k \tag{3.64}$$

$$\dot{\varepsilon}_y = \frac{\partial v}{\partial y} = \sum_{k=1}^{k=p}\dot{\lambda}_k\frac{\partial F_k}{\partial \sigma_y} = \sum_{k=1}^{k=p}\dot{\lambda}_k B_k \tag{3.65}$$

$$\dot{\gamma}_{xy} = \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) = \sum_{k=1}^{k=p} \dot{\lambda}_k \frac{\partial F_k}{\partial \tau_{xy}} = \sum_{k=1}^{k=p} \dot{\lambda}_k C_k \tag{3.66}$$

Differentiating Equations 3.57 and 3.58 with respect to the coordinates and then substituting into the equations above, the flow rule constraint for each element may be written as:

$$\sum_{i=1}^{i=3} \frac{\partial N_i}{\partial x} u_i - \sum_{k=1}^{k=p} \dot{\lambda}_k A_k = 0 \tag{3.67}$$

$$\sum_{i=1}^{i=3} \frac{\partial N_i}{\partial x} v_i - \sum_{k=1}^{k=p} \dot{\lambda}_k B_k = 0 \tag{3.68}$$

$$\sum_{i=1}^{i=3} \frac{\partial N_i}{\partial x} v_i + \sum_{i=1}^{i=3} \frac{\partial N_i}{\partial x} u_i - \sum_{k=1}^{k=p} \dot{\lambda}_k C_k = 0 \tag{3.69}$$

Substituting Equations 3.24 to 3.26 into 3.67 to 3.69, the matrix form of these flow rule constraints is:

$$\frac{1}{2A}\begin{bmatrix} y_{23} & 0 & y_{31} & 0 & y_{12} & 0 \\ 0 & x_{32} & 0 & x_{13} & 0 & x_{21} \\ x_{32} & y_{23} & x_{13} & y_{31} & x_{21} & y_{12} \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \end{bmatrix} - \begin{bmatrix} A_1 & A_2 & A_3 & \cdots & A_k & \ldots & A_p \\ B_1 & B_2 & B_3 & \ldots & B_k & \ldots & B_p \\ C_1 & C_2 & C_3 & \ldots & C_k & \ldots & C_p \end{bmatrix} \cdot \begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \\ \vdots \\ \dot{\lambda}_k \\ \vdots \\ \dot{\lambda}_p \end{bmatrix} = 0 \tag{3.70}$$

which can be written as:

$$\mathbf{a}_{11}\,\mathbf{x}_1 - \mathbf{a}_{12}\,\mathbf{x}_2 = \mathbf{0} \tag{3.71}$$

There are p number of inequality constraints on the plastic multiplier of the form:

$$\mathbf{x}_2 \geq \mathbf{0} \tag{3.72}$$

The flow rule constraints defined by Equation 3.70 must be satisfied by every triangle in the mesh.

3.5.2 Constraints from Plastic Shearing in Discontinuities

A typical velocity discontinuity is shown in Figure 3.7. The discontinuity occurs at the common edge between two adjacent triangles, defined by nodal pairs (1, 2) and (3, 4) and is of zero thickness. To be kinematically admissible, the normal and tangential velocity jump across the discontinuity must satisfy the flow rule, which for a Mohr-Coulomb yield criterion is of the form:

$$\Delta v = |\Delta u| \tan\phi \tag{3.73}$$

where Δv is the normal velocity jump and Δu is the tangential velocity jump. The absolute value on the right hand side of the equation is necessary because, for a non-zero friction angle, dilation occurs regardless of the sign of tangential shearing. For any pair of nodes along the discontinuity (i, j), the tangential and normal velocity jumps are defined in terms of the Cartesian nodal velocities by:

$$\Delta u_{ij} = \left(u_j - u_i\right)\cos\theta + \left(v_j - v_i\right)\sin\theta \tag{3.74}$$

$$\Delta v_{ij} = \left(u_j - u_i\right)\sin\theta + \left(v_j - v_i\right)\cos\theta \tag{3.75}$$

where θ is the angle of discontinuity to the x-axis. The major drawback of this is absolute sign prevents the upper bound formulation being cast as a standard linear programming problem.

To solve this problem, Sloan and Kleeman (1995) proposed each nodal pair (i, j) along a discontinuity be associated with two non-negative variables u_{ij}^{+} and u_{ij}^{-} as

shown in Figure 3.8, which gives rise to two additional unknowns. The tangential velocity jump at each nodal pair, Δu_{ij}, is defined as the difference between these quantities according to:

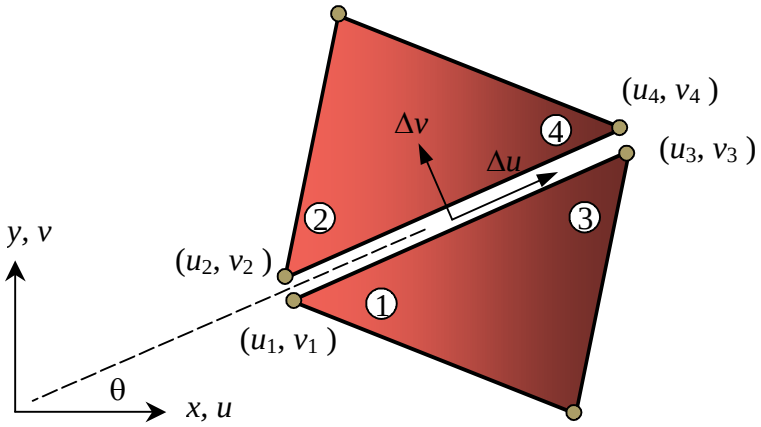


Figure 3.7 Velocity discontinuity geometry.

To solve this problem, Sloan and Kleeman (1995) proposed each nodal pair (i, j) along a discontinuity be associated with two non-negative variables u_{ij}^{+} and u_{ij}^{-} as shown in Figure 3.8, which gives rise to two additional unknowns. The tangential velocity jump at each nodal pair, Δu_{ij}, is defined as the difference between these quantities according to:

$$\Delta u_{ij} = u_{ij}^{+} - u_{ij}^{-} \tag{3.76}$$

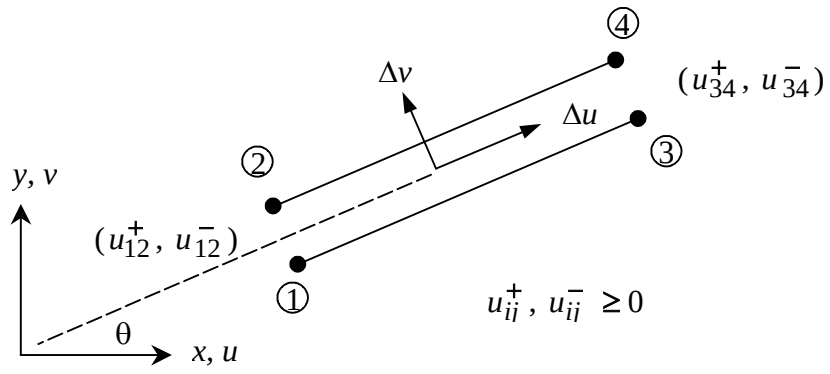


Figure 3.8 Velocity discontinuity variables.

with constraints: $u_{ij}^{+} \geq 0 \qquad u_{ij}^{-} \geq 0$ (3.77)

and: $u_{ij}^{+} - u_{ij}^{-} = (u_j - u_i)\cos\theta + (v_j - v_i)\sin\theta$ (3.78)

In order to be compatible with the velocity expansions in the triangles, the quantities u^{+}, u^{-} and Δu are assumed to vary linearly along the discontinuity according to:

$$u^{+} = u_{12}^{+} + \frac{\xi}{l}\left(u_{34}^{+} - u_{12}^{+}\right) \tag{3.79}$$

$$u^{-} = u_{12}^{-} + \frac{\xi}{l}\left(u_{34}^{-} - u_{12}^{-}\right) \tag{3.80}$$

$$\Delta u = u^{+} - u^{-} = \left(u_{12}^{+} - u_{12}^{-}\right) + \frac{\xi}{l}\left(u_{34}^{+} - u_{34}^{-} - u_{12}^{+} + u_{12}^{-}\right) \tag{3.81}$$

where l is the length of the discontinuity and $0 \leq \xi \leq l$. As a result of Equations 3.77, 3.79 and 3.80, the quantities u^{+} and u^{-} are always non-negative. The tangential velocity jump Δu, however, is unrestricted and may even change sign at some point along the discontinuity. To avoid the need for absolute value signs in the flow rule relations, $|u^{+} - u^{-}|$ is replaced by $(u^{+} + u^{-})$ in Equation 3.73 so that the normal velocity is given by:

$$\Delta v = \left(u^{+} + u^{-}\right) \tan\phi \tag{3.82}$$

this equality is imposed at both nodal pairs (i, j) along the discontinuity according to:

$$\Delta v_{ij} = \left(u_{ij}^{+} + u_{ij}^{-}\right) \tan\phi \tag{3.83}$$

Substituting for Δv_{ij} using Equation 3.75, the final flow rule constraints that are imposed on each nodal pair become:

$$\left(u_i - u_j\right)\sin\theta + \left(v_i - v_j\right)\cos\theta = \left(u_{ij}^{+} + u_{ij}^{-}\right)\tan\phi \tag{3.84}$$

Equations 3.77, 3.79, 3.80 and 3.84 are enforced on each pair of discontinuity nodes. For the discontinuity of Figure 3.8 these constraints may be written in matrix form as follows:

$$\begin{bmatrix} \mathrm{T}_3 & 0 \\ 0 & \mathrm{T}_3 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ u_4 \\ v_4 \end{bmatrix} - \begin{bmatrix} \mathrm{T}_4 & 0 \\ 0 & \mathrm{T}_4 \end{bmatrix} \cdot \begin{bmatrix} u_{12}^{+} \\ u_{12}^{-} \\ u_{34}^{+} \\ u_{34}^{-} \end{bmatrix} = 0 \tag{3.85}$$

where:

$$\mathrm{T}_3 = \begin{bmatrix} -\cos\theta & -\sin\theta & \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta & -\sin\theta & \cos\theta \end{bmatrix} \tag{3.86}$$

$$\mathrm{T}_4 = \begin{bmatrix} 1 & -1 \\ \tan\phi & \tan\phi \end{bmatrix} \tag{3.87}$$

which can be written in form of:

$$\mathbf{a}_{21}\,\mathbf{x}_1 - \mathbf{a}_{23}\,\mathbf{x}_3 = \mathbf{0} \tag{3.88}$$

3.5.3 Constraints from Velocity Boundary Conditions

To be kinematically admissible, the computed velocity field must satisfy the prescribed boundary conditions. Consider a node *i* on a boundary which is inclined at an angle θ to the *x*-axis. For the general case, where the boundary is subject to a prescribed tangential velocity $\bar{u}$ and a prescribed normal velocity $\bar{v}$, the nodal velocity components (u_i, v_i) must satisfy the equalities:

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \overline{u} \\ \overline{v} \end{bmatrix} \tag{3.89}$$

which may be expressed in matrix form as:

$$\mathbf{a}_{31}\,\mathbf{x}_{1} = \mathbf{b}_{3} \tag{3..90}$$

The above type of velocity boundary condition may be used to define the loading caused by a stiff structure, such as a rigid strip footing or retaining wall. In cases where part of the body is loaded by a uniform normal pressure, such as a flexible strip footing, it is often convenient to impose constraints on the surface normal velocities of the form:

$$\int_S v \, \mathrm{d}S = Q \tag{3.91}$$

where Q is a prescribed flow rate of material across the boundary, S, and is typically set to unity. This type of constraint, when substituted into the power expended by the external loads, permits an applied uniform pressure to be minimised directly. Since the velocities vary linearly, This type of constraint may be expressed in terms of the nodal velocities according to:

$$\frac{1}{2} \sum_{\text{edges}} \left[\left(v_i + v_j \right) \cos\theta_{ij} - \left(u_i + u_j \right) \sin\theta_{ij} \right] l_{ij} = Q \tag{3.92}$$

where l_{ij} and θ_{ij} denote the length and inclination of each segment on S, respectively, and each segment is defined by the end nodes (i, j). This boundary condition may be written in matrix form as follows:

$$\frac{1}{2} \begin{bmatrix} -l_{12}\sin\theta_{12} & l_{12}\cos\theta_{12} & -l_{12}\sin\theta_{12} & l_{12}\cos\theta_{12} & \ldots \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} Q \end{bmatrix} \tag{3.93}$$

which may be expressed in matrix form as:

$$\mathbf{a}_{41}\,\mathbf{x}_{1} = \mathbf{b}_{4} \tag{3.94}$$

Another type of constraint arises when a body is subjected to loading by gravity. In this case, it is sometimes convenient to constrain the velocity field so that:

$$\sum_{\text{triangles}} \iint_{A} v \, \mathrm{d}A = -W \tag{3.95}$$

where W is a prescribed constant which is typically set to unity. This constraint permits the unit weight, γ, to be minimised directly when the power expended by the external loads is equated to the internal power dissipation and is useful, such as when analysing the behaviour of slopes. Noting that:

$$\int_{A} v \, \mathrm{d}A = \frac{A}{3}\left(v_1 + v_2 + v_3\right) \tag{3.96}$$

for each triangle, the constraint of Equation 3.95 may be written as:

$$\frac{1}{3}\begin{bmatrix} 0 & A_1 & 0 & A_1 & 0 & A_1 & \cdots \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} -W \end{bmatrix} \tag{3.97}$$

which may be expressed in matrix form as:

$$\mathbf{a}_{51}\,\mathbf{x}_{1} = \mathbf{b}_{5} \tag{3.98}$$

3.5.4 Formulation of Upper Bound Objective Function

A key feature of the formulation is that plastic flow may occur in both the continuum and the velocity discontinuities. The total power dissipated in these modes constitutes

the objective function and is expressed in terms of the unknowns. Within each triangle, the power dissipated by the plastic stresses is given by:

$$P_c = \int_A (\sigma_x \dot{\varepsilon}_x + \sigma_x \dot{\varepsilon}_x + \tau_{xy} \dot{\gamma}_{xy}) \mathrm{d}A \tag{3.99}$$

After substituting Equations 3.64, 3.65 and 3.66 into 3.99 and collecting terms, this dissipation may also be expressed as:

$$P_c = 2A \cos\phi \sum_{k=1}^{k=p} \dot{\lambda}_k \int_A c \, \mathrm{d}A \tag{3.100}$$

If we assume that the cohesion varies linearly throughout the triangle, this integral may be evaluated analytically to give:

$$\boldsymbol{P}_c = \mathbf{c}_2^T \mathbf{x}_2 \tag{3.101}$$

where:

$$\mathrm{c}_2^T = \frac{2}{3} A(c_1 + c_2 + c_3) \cos\phi \, [1 \quad 1 \quad \ldots] \qquad \mathrm{x}_2 = \begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \\ \vdots \\ \dot{\lambda}_p \end{bmatrix} \tag{3.102}$$

where c_1, c_2, c_3 are nodal values of the cohesion. Since the plastic multiplier rates are constrained so that $x_2 \geq 0$, it follows that the power dissipated in each triangle is non-negative.

The power dissipated by plastic shearing along a velocity discontinuity is given by an integral of the form

$$\int_l c|\Delta u| \, \mathrm{d}l \tag{3.103}$$

For the discontinuity shown in Figure 3.8, $(u^+ + u^-)$ is substituted for $|\Delta u|$ in order to preserve the linear character of the objective function according to:

$$P_{\mathrm{d}} = \int_{l} c(u^{+} + u^{-})\mathrm{d}l \tag{3.104}$$

After substitution using Equations 3.79 and 3.80 and integrating, the power dissipated in each discontinuity may be written as:

$$\boldsymbol{P}_{\mathbf{d}} = \mathbf{c}_{\mathbf{3}}^{\mathsf{T}}\,\mathbf{x}_{\mathbf{3}} \tag{3.105}$$

where:

$$\mathrm{c}_{3}^{\mathrm{T}} = l\left[\frac{1}{3}c_1 + \frac{1}{6}c_2 \quad \frac{1}{3}c_1 + \frac{1}{6}c_2 \quad \frac{1}{6}c_1 + \frac{1}{3}c_2 \quad \frac{1}{6}c_1 + \frac{1}{3}c_2\right] \tag{3.106}$$

$$\mathrm{x}_{3} = \begin{bmatrix} u_{12}^{+} \\ u_{12}^{-} \\ u_{34}^{+} \\ u_{34}^{-} \end{bmatrix} \tag{3.107}$$

and the cohesion varies linearly so that c_1 and c_2 are the cohesions at the nodal pairs (1, 2) and (3, 4) respectively. Note that this form of power dissipation is also non-negative, since the discontinuity variables are constrained so that $x_3 \geq 0$.

All the constraint coefficients and objective function coefficients may be assembled to give an upper bound linear programming problem, which is expressed as:

Minimise:	$\mathbf{c}_{\mathbf{2}}^{\mathsf{T}}\,\mathbf{x}_{\mathbf{2}} + \mathbf{c}_{\mathbf{3}}^{\mathsf{T}}\,\mathbf{x}_{\mathbf{3}}$
Subject to:	$\mathbf{a}_{11}\,\mathbf{x}_{1} - \mathbf{a}_{12}\,\mathbf{x}_{2} = \mathbf{0}$
	$\mathbf{a}_{21}\,\mathbf{x}_{1} - \mathbf{a}_{23}\,\mathbf{x}_{3} = \mathbf{0}$
	$\mathbf{a}_{31}\,\mathbf{x}_{1} = \mathbf{b}_{3}$
	$\mathbf{a}_{41}\,\mathbf{x}_{1} = \mathbf{b}_{4}$
	$\mathbf{a}_{51}\,\mathbf{x}_{1} = \mathbf{b}_{5}$
	$\mathbf{x}_{2} \geq \mathbf{0}$
	$\mathbf{x}_{3} \geq \mathbf{0}$

(3.108)

where $\mathbf{x}_1$ represents vector of nodal velocity, whilst $\mathbf{x}_2$ represents vector of element plastic multiplier rates, and $\mathbf{x}_3$ represents vector of discontinuity parameters.

3.6 NONLINEAR FORMULATION OF LOWER BOUND AND UPPER BOUND THEOREM

The preceding section was concerned with the formulation of the linear implementation of the lower and upper bound theorems introduced by Sloan (1988) and Sloan and Kleeman (1995), respectively. Although both lower bound and upper bound finite element formulations based on linear programming have proved successful for the solution of two-dimensional stability problems, they are unsuitable for three-dimensional geometries, because the linearisation of the three-dimensional yield surface inevitably generates a large number of linear inequalities. This in turn, leads to slow solution times for the linear programming solver that conducts a vertex-to-vertex search. Alternatively, one could formulate a lower bound and upper bound scheme using the combination of linear finite element and the nonlinear programming procedure (Lyamin and Sloan, 2002a, 2002b).

The Mohr-Coulomb yield criterion presents a number of computational difficulties due to the gradient discontinuities at the tip and edges of the hexagonal yield surface pyramid. To resolve this difficulty, Lyamin and Sloan (2002a, 2002b) have adopted a smooth approximation of the original yield surface to remove all gradient singularities from the Mohr-Coulomb yield criterion. A convenient hyperbolic approximation for smoothing the tip and corners of the Mohr-Coulomb model, which requires just two parameters, has been given by Abbo and Sloan (1995). A plot of the hyperbolic approximation to the Mohr Coulomb yield function in the meridional plane is shown in Figure 3.9.

This new approach uses the yield criterion in its native non-linear form, and both three-dimensional stress and velocity fields present no particular difficulty apart from the complex geometry. For the lower bound formulation, the objective function and the equality constraints are linear, while the yield inequalities are non-linear. The problem of finding a statically admissible stress field which maximizes the collapse load may be expressed as:

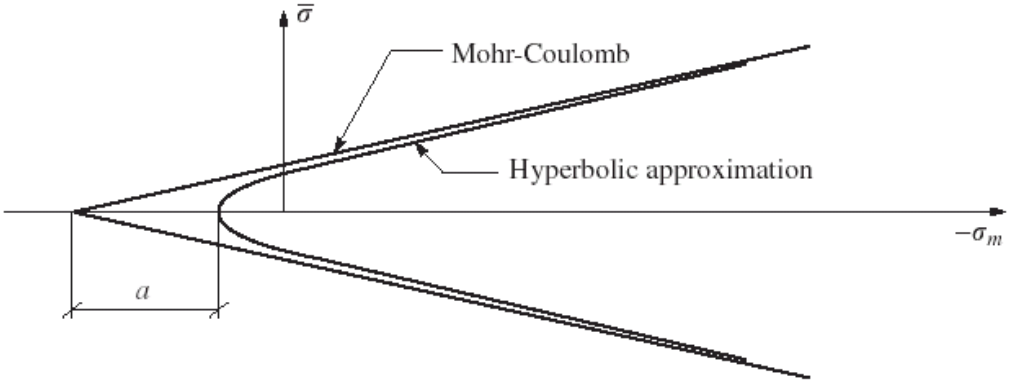


Figure 3.9 Hyperbolic approximation to Mohr-Coulomb yield function. (*After Lyamin and Sloan, 2002*).

$$\begin{aligned} &\textit{Maximise:} && \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ &\textit{Subject to:} && \mathbf{A}\mathbf{x} = \mathbf{b} \\ & && f_i(\mathbf{x}) \leq 0,\ j \in J \\ & && \mathbf{x} \in \mathrm{R}^n \end{aligned} \tag{3.109}$$

where **c** is an *n*-vector of objective function coefficients, **A** is an $m \times n$ matrix of equality constraint coefficients, $f_i(\mathbf{x})$ are yield functions and other inequality constraints, and **x** is a the *n*-vector of problem unknowns.

When formulating the upper bound non-linear programming problem, for a dilational yield criterion whose shape is non-linear in the meridional plane, the flow rule constraints are applied using a local Mohr-Coulomb approximation to the yield surface. The problem of finding a kinematically admissible velocity field, which minimises the internal power dissipation, may be written as:

$$\begin{aligned} &\textit{Minimise:} && Q = \sigma^{\mathrm{T}}\mathbf{B}\mathbf{u} + \mathbf{c}_u^{\mathrm{T}}\mathbf{u} + \mathbf{c}_v^{\mathrm{T}}\mathbf{v} \quad \text{on } \mathbf{(u, v)} \\ &\textit{Subject to:} && \mathbf{A}_u\mathbf{u} + \mathbf{A}_v\mathbf{v} = \mathbf{b} \\ & && \mathbf{B}\mathbf{u} = \sum_{j \in J_\sigma} \lambda_j \nabla f_j(\sigma) \\ & && \lambda_j \geq 0,\ \ j \in J_\sigma \end{aligned} \tag{3.110}$$

$$\lambda_j f_j(\sigma) = 0, \quad j \in J_\sigma$$

$$f_j(\sigma) \leq 0, \quad j \in J_\sigma$$

$$\mathbf{v} \geq 0$$

$$\mathbf{u} \in \mathrm{R}^{n_u}, \mathbf{v} \in \mathrm{R}^{n_v}, \sigma \in \mathrm{R}^{n_\sigma}, \lambda \in \mathrm{R}^{n_\lambda}$$

where **B** is a global compatibility matrix, $\mathbf{c}_u$ is a vector of objective function coefficients for the velocities, $\mathbf{c}_v$ is a vector of objective function coefficients for the discontinuity variables, $\mathbf{A}_u$ is a matrix of equality constraint coefficients for the velocities, $\mathbf{A}_v$ is a matrix of equality constraints coefficients for the discontinuity variables, $f_j(\sigma)$ are yield functions, λ_j are non-negative multipliers, and **u**, **v** and $\boldsymbol{\sigma}$ are problem unknowns.

To solve these resulting systems, one transforms Equations 3.109 and 3.110 into a set of Kuhn-Tucker optimality conditions, and the rapid convergent two-stage quasi-Newton method is needed to obtain the optimum solutions (Lyamin and Sloan, 2002a, 2002b). The key advantage of this strategy is that its iteration count is largely unaffected by the mesh size, and the resulting formulation is many times faster than an equivalent linear programming formulation. Comparisons presented by Lyamin and Sloan (2002a, 2002b) suggested that this new strategy offers up to an impressive 155-fold reduction in the CPU time for the analysis, which used the mesh with a large number of elements and discontinuities.

3.7 DISPLACEMENT FINITE ELEMENT METHOD

The displacement finite element method (DFEM) is the most commonly used and most versatile method for constructing approximate solutions to boundary-value problem. The benefits and limitations of the DFEM were summarised by Carter et al. (2000). The DFEM method involves discretising the region of interest into a finite number of sub-regions, and establishing governing differential equations for each element in the form of matrix equations. A variety of concepts can be used to construct an approximation of the solution over the collection of finite elements. Detailed descriptions of DFEM are given in a large number of textbooks (e.g. Smith

and Griffiths, 1998), and, due to the proliferation of this method in geotechnical practice, there is no need to duplicate such detail in this thesis.

In order to provide reliable and accurate solutions to different geotechnical problems, a complete constitutive model for geomaterials is required to provide the stress-strain behaviour prior to failure. Linear elastic and elastic-plastic are the most commonly used constitutive models due to their simplicity.

In order to solve the problem that arises from the additional kinematic constraints imposed on the system by the specified volumetric strains associated with plastic flow, Sloan pioneered the use of high order elements (Sloan, 1981; Sloan and Randolph, 1982). He proposed the use of the 15-noded triangular element for accurate collapse load predictions under conditions of plane strain and axisymmetry. The 15-noded triangular element behaves well for both purely cohesive (incompressible) and cohesive-frictional (dilatational) materials and is relatively simple to implement. Smith and Griffiths (1998) suggested that the eight-noded quadrilateral element, together with reduced integration (four Gauss points per element) is able to compute collapse loads accurately (Zienkiewicz, 1991; Griffiths, 1982b). In order to reduce “locking” problems, they recommended that the order of integration of the volumetric components could also be reduced further.

Although the speed and storage capabilities of computers have dramatically increased over the last decades, many practical problems are still only tractable with severe simplifications. Moreover, there is also a need to examine the quality of the solution obtained to ensure its reliability. Hence, one is interested to use as few degrees of freedom as possible in the problem, while on the other hand the obtained solution should be as accurate as possible. Much research has focused on the development of efficient matrix storage strategies, and summaries of those that are commonly used in practice are given in Table 3.2.

Depending on the computing resources available, different matrix storage strategies are applied. Element-by-element storage is the most preferred method used when performing large scale, three-dimensional modelling due to its low memory requirements. The implementation of element-by-element storage requires the

collection of element matrices to be solved iteratively using numerical technique such as conjugate gradient method (Smith and Griffiths, 1998). Another benefit of implementing element-by-element storage is that it allows for parallel computation as the element level matrices can be solved in parallel without affecting one another.

The treatment to this point has involved homogeneous soil layers; that is, where each adjacent element has identical soil properties. The following sections examine the theory of random fields and the application of heterogenous soil profiles in reliability studies.

Table 3.2 Summaries of different matrix storage strategies.

Skyline storage	• Store all coefficients within a skyline; • Requires renumbering to minimise skyline; • Efficient for direct solvers on shared-memory machines.
Sparse storage	• Stores non-zero coefficients only; • Predominantly indirect access patterns; • Efficient for iterative methods on distributed memory architectures.
Element by element storage	• Retains element-level matrices, no global assembly; • Efficient for iterative methods on distributed memory; • The main advantages of element-by-element solvers are significant saving in numerical operations, computer storage and potential for parallel implementation; • Element matrices can be recalculated as needed, minimal storage, but slower.

3.8 HETEROGENEOUS SOILS

The enterprises of structural and mechanical engineers, and geotechnical engineers are significantly different. While structural and mechanical engineers deal with manufactured materials, geotechnical engineers deal almost exclusively with natural materials. Soils are formed from parent materials, which are commonly heterogenous. However, soil properties do not really vary *randomly*, but exhibit some degree of continuity between adjacent locations. This is due to the manner in which

geotechnical materials are formed and the continuous complex and varied environment processes, which work to alter them (Spry et al., 1988).

3.8.1 Random Field Theory

A mathematical model can be used to describe such a random process. Consider any random process, $Z(x)$ (physical or mechanical in nature), defined along the points x of a given domain. This random process is classically described in the following form:

$$Z(x) = m(x) + \varepsilon(x) \tag{3.111}$$

where $Z(x)$ is a random field process; $m(x)$ is a deterministic function describing the mean of the soil property at x; and $\varepsilon(x)$ is the random residual or *white noise* (Fenton, 1999).

In this model, $m(x)$ represents a linear or quadratic trend of the random field process, $Z(x)$, and $\varepsilon(x)$ represents the random fluctuation of this process about the trend (Fenton, 1999). This model can represent any particular soil property (e.g. undrained shear strength) that fluctuates along a vertical or horizontal line within a soil mass. However, spatial variability analyses are not conceptually limited to horizontal or vertical directions, though these orientations are the most significant and frequently investigated in practice. Jaksa (1995) found that the cone tip resistance, q_c, data of Keswick Clay exhibits an underlying quadratic trend pattern that was consistent with Keswick Clay being overconsolidated as a result of desiccation. Noted that the trends can be as complex as desired by the user and are not necessarily limited to first- or second-degree polynomials. However, Fenton (1999) suggested that the trend component might be a part of a fractal process and it should be regarded as a part of the overall uncertainty, and therefore should not be removed from the stochastic characterisation, in order to obtain global results.

Given sufficient data gathered from the field, one could accurately characterise the spatial variability of soil properties using statistical formulations. The main elements for characterising the spatial variability of soil properties include (Vanmarcke, 1977a):

- Classical statistical properties of the soil (i.e. mean, coefficient of variation (*COV*) and distribution of data); and
- Spatial correlation structures of soil (which comprise the covariance relationship) and scale of fluctuation (which describe the semi-continuity and randomness of the soil property in space).

If an infinite amount of data were available, one could gain only a pseudo-complete spatial description of soil variability in any direction, at least because measurements are always inaccurate and imprecise to some degree (Baecher, 1982). As data are always limited in number, it becomes convenient to model soil variability as a random variable. Structures explanations of the statistical techniques used for the investigation of spatial variability are presented by Priestley (1981) and Baecher and Christian (2003).

Vanmarcke (1978) suggests that the magnitude of spatial correlation depends from the method and scale of investigation. Vanmarcke (1978) illustrates the multiplicity of scales of investigation which are possible in geotechnical engineering: (a) soil particles; (b) laboratory specimen; (c) vertical sampling; (d) lateral distance between borings; (e) horizontal intervals at regional scale, e.g. measured along the centreline of long linear facilities. Distinct measurement techniques are generally employed for the various scales. The dependence of the correlation structure on the scale of investigation had led a number of researchers (e.g. Agterberg, 1970; Campanella et al., 1987; Fenton, 1999a, 1999b; Fenton and Vanmarcke, 2003; Jaksa, 1995; Kulatilake and Um, 2003) to address autocorrelation in soils using fractal models, which assume an infinite correlation structure and allow to directly address the dependence of the correlation structure from the sampling domain.

3.8.2 Classical Statistical Properties

Several classical statistical characteristics of soils, such as the mean, *COV* and probability distribution, have been published by a number of researchers (e.g. Lumb, 1966, 1970; Li and White, 1987; Spry et al., 1988; Jaksa, 1995; Jaksa et al., 1996, 1999; Phoon and Kulhawy, 1999a, 1999b). It is suggested that, generally, the strength parameters usually exhibit high variability, in terms of high *COV* (Spry et al., 1988).

The statistical distributions that are usually adopted for geotechnical properties by different authors are the normal, lognormal and beta distributions (Lee et al., 1983; Paice et al., 1996; Kaggwa, 2000; Griffiths and Fenton, 2001; Griffith et al., 2002; Fenton and Griffith, 2002, 2003). It is found that these statistical distributions are site and parameter specific and there is no generic probability distribution pattern for all soil properties.

Second-moment statistics (i.e. mean and standard deviation) alone are unable to describe the spatial variation of soil properties, whether measured in the laboratory or in-situ. El-Ramly et al. (2002) have demonstrated that two sets of measurements may have similar second-moment statistics and statistical distributions, but could display substantial differences in spatial distribution.

3.8.3 Spatial Correlation

The *auto-correlation function* (ACF) is a measure of similarity between two or more points in a data series, which can be defined as:

$$B(h) = \mathrm{E}[\, Z(x)\, Z(x+h)] - \overline{Z}^2 \qquad (3.112)$$

where $B(h)$ is the corresponding value of the autocorrelation function at a separation distance h; $Z(x)$ and $Z(x+h)$ are pairs of data separated by distance h; $E[\ldots]$ is the expected value operator; and $\overline{Z}$ is the mean value of data series Z. The autocorrelation function is used widely throughout time series analysis literature, and it enables the characteristic of the time series to be determined. For example, a slowly decaying autocorrelation function suggests long-term dependence, whilst a rapidly decaying function suggests short-term dependence (Chatfield, 1975; Hyndman, 1990).

Vanmarcke (1977a, 1983) suggested that the spatial variability of geotechnical materials may be characterised stochastically by the use of three parameters; namely, the mean, μ, the standard deviation, σ, and the scale of fluctuation, δ. The scale of fluctuation describes the spatial continuity of the random field; that is, values at

adjacent locations are more correlated than those separated by larger distances. In other words, the scale of fluctuation quantifies the spatial extension of strong autocorrelation. The spatial continuity of a random field is parameterised by the autocorrelation function. Vanmarcke (1983) proposed that δ may be evaluated by fitting a model ACF to the sample autocorrelation. Some typically used model autocorrelation functions and their relationship to δ, are shown in Table 3.3.

Table 3.3 Scale of fluctuation with respect to theoretical autocorrelation functions. (*After Vanmarcke, 1977a, 1983*)

Model No.	Autocorrelation Function	Scale of Fluctuation, δ
1	$\rho_h = e^{-\lvert h\rvert / a}$	$2a$
2	$\rho_h = e^{-(\lvert h\rvert / b)^2}$	$(\pi^{1/2})\, b$
3	$\rho_h = e^{-\lvert h\rvert / c}\left(1+\frac{\lvert h\rvert}{c}\right)$	$4c$

3.8.4 Local Average Subdivision (LAS)

Stochastic finite element modelling and analysis involves finite element modelling incorporating random field simulation. Whilst there are several random field simulation techniques available and discussed in the literature, one method, namely local average subdivision (LAS), offers a fast and reliable simulation technique that, as a results, is employed in the present work.

The local average subdivision (LAS) approach was introduced by Fenton and Vanmarcke (1990). This method offers a fast and accurate means to generate realisations of homogeneous Gaussian random fields in one, two or three dimensions. The first of its type appeared as *stochastic subdivision* algorithms described by Carpenter (1980) and Fournier et al. (1982) in computer graphics synthesis. It was limited, however, in that it was only able to model power spectra having a form of $w^{-\beta}$ with a single parameter β. It was later generalised by Lewis (1987) eliminating this limitation, and extending its capabilities to modelling fields with non-fractal spectral and directional or oscillatory characteristics.

Fenton and Vanmarcke (1990) described a local averaging process by means of LAS, which essentially proceeds in a top-down recursive fashion, as illustrated in Figure 3.10. In the first Stage 0, a global average is generated for the process. In the next stage (Stage 1), the domain is subdivided into two sub-domains whose local averages must in turn be the average of the global value, i.e.

$$Z_1^0 = \frac{1}{2} Z_1^1 + \frac{1}{2} Z_2^1 \tag{3.113}$$

NOTE:
This figure is included on page 76 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.10 Top down approach to the LAS construction. (*After Fenton and Vanmarcke, 1990*).

Fenton and Vanmarcke (1990) defined the term *parent cell* as the previous stage cell being subdivided and, *within cell* meaning the region defined by the parent cell. Further subdividing each of the *parent cells* into two equal size sub-domains then generates the subsequent stages. Normally distributed values are generated for the resulting two sub-domains while preserving upward averaging. These values are generated such that they must exhibit the correct mean and variance according to local averaging theory. In addition, these generated values must be appropriately correlated with one another, as well as correlated with their neighbouring values.

In order to determine the mean and variance of Stage 0, Z_1^0, consider a continuous stationary scalar random function $Z(t)$ in one dimension, and a domain of interest $(0, D]$. At this stage, two points are noted: first, the implementation of this method is strictly applied to stationary processes only, which are fully described by their second order statistics (mean, variance, and autocorrelation function). Second, the LAS

method depends on the size of the domain, over which local averaging is performed, being known.

Z_1^0 is a random variable whose statistics can be expressed by using stationary theory and the fact that $B(\tau)$, the covariance function of $Z(t)$, is an even function of lag τ. Without loss of generality, $E[Z]$ will henceforth be taken as zero. It will give sufficient information to generate a realization of Z_1^0 at stage 0 using Equations 3.114 and 3.115, if $Z(t)$ is a Gaussian random process (Fenton, 1990).

$$E[Z_1^0] = E(z) \tag{3.114}$$

$$E[(Z_1^0)^2] = E^2[Z] + \left(\frac{2}{D^2}\right)\int_0^D (D-\tau)B(\tau)\,d\tau \tag{3.115}$$

A complete probability distribution function for Z_1^0 must be determined if $Z(t)$ is not a Gaussian random process, and a realization of Z_1^0 is generated according to such a distribution (Fenton and Vanmarcke, 1990). With reference to Figure 3.11, the determination of Stage $i + 1$ values, given Stage $i + 1$ is obtained by estimating Z_{2j}^{i+1} with the addition of a zero mean discrete white noise $c^{i+1}W_j^{i+1}$, shown in Equation 3.116 and M_{2j}^{i+1} can be determined by Equation 3.117, shown as follows:

$$Z_{2j}^{i+1} = M_{2j}^{i+1} + c^{i+1}W_j^{i+1} \tag{3.116}$$

$$M_{2j}^{i+1} = \sum_{k=j-n}^{j+n} a_{k-j}^{i} Z_k^i \tag{3.117}$$

Fenton (1990) showed that, for $i = 0, 1, 2, \ldots, L$ and $\ell = -n, \ldots, n$, the set of $\{a_\ell^i\}$ and c^{i+1} can be determined by using Equations 3.118 and 3.119 respectively. The adjacent cell Z_{2j-1}^{i+1} is determined using Equation 3.120, to ensure the upwards averaging is preserved.

$$E\left[Z_{2j}^{i+1} Z_m^i\right] = \sum_{k=j-n}^{j+n} a_{k-j}^i E\left[Z_k^{i+1} Z_m^i\right] \tag{3.118}$$

$$\left(c^{i+1}\right)^2 = E\left[\left(Z_{2j}^{i+1}\right)^2\right] - \sum_{k=j-n}^{j+n} a_{k-j}^i E\left[Z_{2j}^{i+1} Z_k^i\right] \tag{3.119}$$

$$Z_{2j-1}^{i+1} = 2Z_j^i - Z_{2j}^{i+1} \tag{3.120}$$

NOTE:
This figure is included on page 78 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.11 1-D LAS indexing schemes for Stage *i* (top) and Stage *i* + 1 (bottom). (*After Fenton, 1990*).

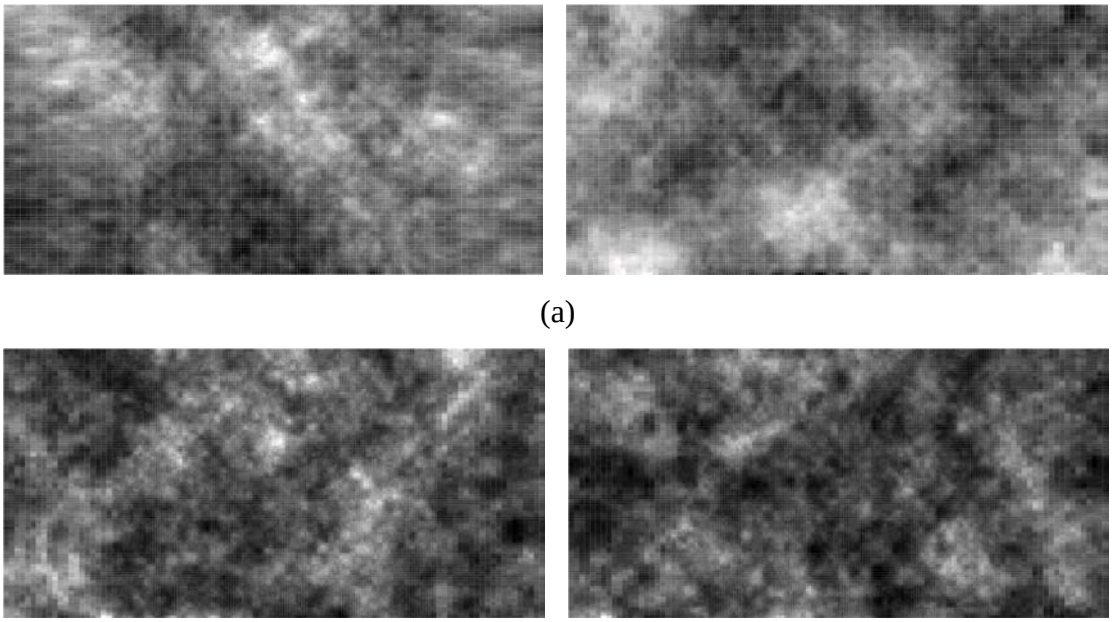
Fenton (1990) employed several theoretical exponentially decaying correlation functions in LAS, one of which is used in this study is called Markovian spatial correlation function and is given by:

$$\rho(\tau) = \exp\left(\frac{-2|\tau|}{\theta}\right) \tag{3.121}$$

where $\rho(.)$ is the correlation coefficient between two points separated by vector $\tau = \{\tau_x, \tau_y\}$ and θ is the correlation length, which defined as, loosely speaking, the distance over which soil properties are significantly correlated (when the $|\tau|$ is greater than θ, the correlation coefficient between two points is less than 14%). This correlation length governs the correlation function decay rate and it will be used throughout this study.

The major drawback of LAS, as indicated by Fenton (1994), is that it is incapable of directly producing an anisotropic field due to its same cross-cell approximation. The

field covariance always tends towards isotropy with a scale equal to the minimum of the anisotropic scales. However, Fenton (1994) suggested that an ellipsoidal anisotropic random field could be produced from an isotropic random field by stretching the coordinate axes. Regardless of its disadvantages, Fenton and Vanmarcke (1990) concluded that LAS is a fast and sufficiently accurate method for generating realisations of random fields and demonstrates accurate covariance at any given resolution. In addition, LAS is suitable for generating random fields for finite element analysis, since the soil medium is represented in element form rather than point form (Fenton and Vanmarcke, 1990). Accordingly, the LAS method will be used in subsequent soil profile simulations. Examples of random fields generated by LAS using two different correlation distances are illustrated in Figure 3.12. Noted that the sizes of the generated random fields and the correlation distances used in this chapter are dimensionless.



(a)

(b)

Figure 3.12 Examples of plausible random fields.
(a) $\theta = 2.0$, (b) $\theta = 64.0$

3.8.5 Applications of Random Field

There has been a steady stream of publications on the applications of reliability methods to geotechnical engineering over the last thirty years or so. The presence of

heterogeneity in properties of natural soil masses and its considerable effect on design performance has been extensively addressed by several researchers (e.g. Vanmarcke, 1977b; Righetti and Harrop-Williams, 1988; Paice et al., 1996; Kaggwa, 2000; Griffiths and Fenton, 2001; Griffith et al., 2002; Fenton and Griffith, 2002, 2003). With the aid of the high speed and powerful computers, the soil medium in finite element analysis can be modelled as a heterogeneous random field by using Vanmarcke's variance function (Harrop-Williams, 1988), the local average subdivision technique (LAS) (e.g. Paice et al., 1996, Fenton et al., 2003, Fenton and Griffiths, 2003) or Monte Carlo simulation. The reliability analyses have been performed to obtain the reliability index and the probability of failure as a function of various coefficients of variation and, either scales of fluctuation or correlation lengths (if LAS is used).

The effect of random elastic moduli on strip foundation settlement was studied by Paice et al. (1996). In order to avoid negative values of Young's modulus, E, in random field simulation, a lognormal probability distribution of E was assumed. There are some limitations of this study, such as linear-elastic soil behaviour, only an isotopic simple-exponential correlation structure was considered, and out-of-plane variation of E was not considered. Despite these limitations, it was shown that the variation of elastic modulus has a significant influence on footing settlement.

Fenton et al. (2003) studied the reliability of Janbu's settlement formulation in the prediction of strip footing settlement on spatially random soil. The settlement of a strip footing was determined by simple linear-elastic, two-dimensional finite element analysis with a stochastic input of the elastic modulus of the soil. The elastic modulus field, again was assumed to follow a lognormal probability distribution and Markovian correlation function, and the random field was simulated using the LAS method. Simulated site investigations were carried out at a limited number of locations on simulated soil profiles, in order to determine the representative value of E to compute the required footing using Janbu's equation. The finite element method was then used to calculate the actual settlement of the designed footing on a particular realization of the elastic modulus field. Monte Carlo simulation was used to assess the reliability of Janbu's settlement formulation with respect to the soil's variance and correlation length. Fenton et al. (2003) concluded that, if the basic statistical data (i.e.

mean, variance and correlation length) are known, Janbu's settlement formulation is reasonably reliable. A framework for assessing the reliability of footing design was proposed by Fenton et al. (2003).

3.8.6 Soil delineation

Random fields are simulated based on an assumption of strict stationarity, that is the mean and standard deviation of the stochastic process or time series are constant over time or position. However, natural soil profiles usually consist of multi-layers, with each layer having different spatial statistics (probability distribution and scale of fluctuation) of soil properties. The significant variation of spatial statistics across a site can violate the assumption of stationarity and can lead to biased estimation. There are two traditional approaches often taken to address such a limitation (Wingle, 1997). One divides the site into a number of different zones and describes the spatial statistic for each zone, simulate or estimate each zone, and merge them together. The other assumes a pooled spatial model that reflects the mean behaviour of the entire site and the problem can be controlled with the local stationarity of the neighbouring data samples (Isaaks and Srivastava, 1989). The former approach requires effort and resources and is not equally suitable for sites with gradual transitions.

An alternative approach presented by Wingle (1997), called *zonal kriging*, has the advantage of describing and defining the inter-zone relationships. These inter-zonal relationships define how the nearest neighbouring data points in one zone influence the simulated cell located in another zone. This approach produces realisations of non-stationary random fields that more accurately represent reality (Wingle, 1997). However, this approach has a few limitations, such as, this approach is only applicable to *simple* and *ordinary* kriging, which is known as a linear unbiased estimator of a spatial process based on a large body of stochastic data modelling methodology known as *geostatistics* proposed by Matheron (1970), and therefore it is only applicable to random field simulation (i.e. *sequential indicator simulation*) by geostatistics.

Due to the limitation of the LAS method, which is incapable of simulating non-stationary random fields, the first aforementioned approach will be adopted. The soil

profiles will be divided into a number of layers and each layer is then simulated independently. The resulting random fields will be merged based on the location of the boundary and thickness of the layers. Such an approach requires prior and post processing and is limited to simulating a multi-layered soil profile with abrupt boundaries.

3.9 ARTIFICIAL NEURAL NETWORKS

The concept of artificial neural networks (ANNs) was first introduced by McCulloch and Pitts in 1943. The basic idea of artificial neural networks (ANNs) is to develop mathematical algorithms whose purpose is to resemble closely the operation of the human brain system or natural neural networks (NNNs) (McCulloch and Pitts, 1943). ANNs may be considered as a new age, powerful tools in the field predicting and forecasting. Recently, ANNs have been applied successfully to a wide range of areas including classification, estimation, prediction and function synthesis (Moselhi et al., 1992). The applications of ANNs in engineering fields, which include geotechnical engineering, have been reviewed extensively by Shahin et al. (2001).

ANNs can be trained by learning from examples or past experiences, which are a set of input variables and corresponding outputs. ANNs are used to determine the inter-variable rules that govern the relationship between variables. Therefore, ANNs are suitable for modelling complex problems where the relationship between the variables is unknown (Hubick, 1992) and non-linearity is presented (Maier, 1995). The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical.

The aim of this section is to provide some background information and important features associated with ANNs. A brief description of natural neural networks (NNNs) will be presented, followed by the structure and operation of ANNs. There are different ANN structures, however, in this research, back-propagation multilayer perceptrons will be employed and therefore, will be discussed in greater detail.

3.9.1 Natural Neural Networks (NNNs)

The structure and operation of natural neural networks (NNNs), which is illustrated in Figure 3.13, have been discussed by several authors (e.g. Hertz at al., 1991; Zurada, 1992; Masters, 1993; Fausett, 1994). In brief, NNNs consist of, billions in number, densely interconnected nerve cells, which are found in the brain, called neurons. Each neuron receives combined signals from many other neurons through synaptic gaps via dendrites. The dendrites collect the incoming signals and send them to the soma, which is the cell body of the neuron. The soma sums the incoming signals and the neuron is activated if the charge of these signals is strong enough. The signal is then transmitted to neighbouring neurons through an output structure called an axon. The axon divides and connects to dendrites of neighbouring neurons through junctions called synapses. The signals that the neural networks receive, process and transmit are electrochemical, which means electronic impulses are transmitted by means of a chemical process (Fausett, 1994). Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

3.9.2 Artificial Neural Networks (ANNs)

A comprehensive description of the structure and operation of ANNs have been discussed and given by many authors (e.g. Hecht-Nielsen, 1990; Maren et al., 1990; Zurada, 1992; Fausett, 1994; Ripley, 1996). Basically, artificial neural networks or ANNs are introduced to imitate the behaviour of NNNs. ANNs are typically composed of interconnected processing units or nodes, which serve as model neurons. A graphical representation of an ANN is shown in Figure 3.14. This model is a generalization known as multi-layered perceptrons (MLPs). The processing units are usually arranged in layers, namely the input layer, hidden layers and output layer. Hidden layers are intermediate layers between the input and output layer and are used to model relationship between input and output layers. Each processing unit in a specified layer is connected to other processing units in other layers.

The function of the synapse is modelled by an adjustable weight. The weight represents the strength of the synapse of the neuron in NNNs and it determines the strength of the connection between two processing units. A zero weight implies no

connection between two processing units, whilst a positive weight represents an excitatory connection, and a negative weight refers to an inhibitory connection.

NOTE:
This figure is included on page 84 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.13 Typical structure of biological neuron (*After Fausett, 1994*)

Each input (x_i) from each processing unit in the previous layer is multiplied by an adjustable weight (w_{ji}). In each processing unit, the weighted inputs ($w_{ji}\ x_i$) are summed and a bias (θ_j) is then added or subtracted. Such combined input (I_j) is then passed through a nonlinear input-output or transfer function [$f(\ )$] (e.g. sigmoidal or tanh) to produce an output. The output from one processing unit serves as an input for the processing units in the following layer.

Summation: $$I_j = \Sigma\, w_{ji}\, x_i + \theta_j \tag{3.122}$$

Transfer: $$y_j = f(I_j) \tag{3.123}$$

The learning or training process of ANNs involves adjusting weights such that the error between the measured or historical data and the output from the ANN is minimised. Such a process requires computation of the error derivative of the weights; that is, the change in error with respect to the change in the weights. The most commonly used method for determining such an error derivative is the back propagation algorithm. The algorithm first computes the error function, which is the rate of error changes as the activity of a processing unit is changed. For example, the error function, for node *j*, is given by following equation:

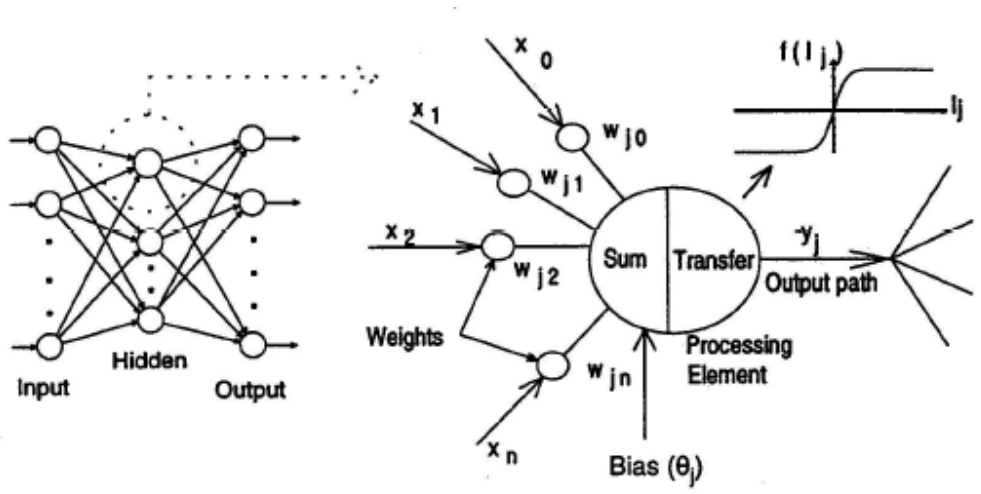


Figure 3.14 Typical structure and operation of ANNs (*After Maier and Dandy, 1998*)

$$E = \frac{1}{2}\sum\left(y_j - d_j\right)^2 \tag{3.124}$$

where:

E = the global error function;

y_j = the predicted output by the network; and

d_j = the historical or measured actual output.

The global error function, E, is minimised by modifying the weights using the gradient descent rule:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} \tag{3.125}$$

where:

Δw_{ji} = weight increment from node i to node j; and

η = learning rate, by which the size of the step taken along the error surface is determined.

The equation can be further refined by the *delta rule* as follows:

$$\Delta w_{ji} = \eta \, \delta_j \, x_i \tag{3.126}$$

where

x_i = input from node i, $i = 0, 1, 2 \ldots, n$; and

δ_j = error value between the predicted and desired output for node j.

The back propagation algorithm is applied to adjust the weights in a backward manner. The weights between the hidden and the output layers are adjusted first, followed by the weights in between the hidden and the input layers. If node j is in the output layer, then δ_j can be written as follows:

$$\delta_j = (y_j - d_j) \, f'(I_j) \tag{3.127}$$

where

$f'(I_j)$ = the derivative of the activation function f with respect to the weighted sum of inputs of node j.

If node j is in the hidden layer, with reference to Figure 3.15, by applying the generalised delta rule proposed by Rumelhart et al. (1986), δ_j can be determined as follows:

$$\delta_j = \left[\sum_{1}^{m} \delta_m w_{mj} \right] f'(I_j) \tag{3.128}$$

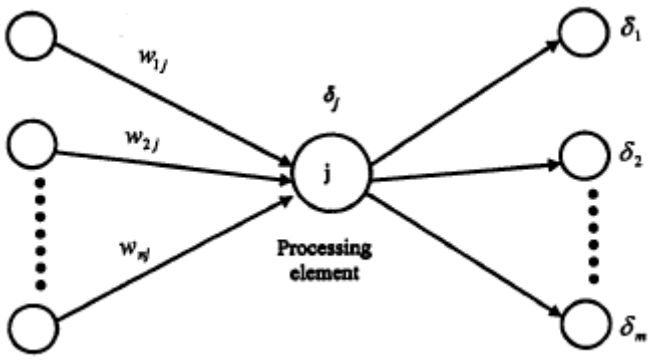


Figure 3.15 Node j in hidden layer.

The weights are adjusted by adding the delta weight, Δw_{ji}, to the corresponding previous weight as follows:

$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji} \tag{3.129}$$

where:

$w_{ji}(n+1)$ = the weight from node i to node j at step n (before adjustment); and

$w_{ji}(n)$ = the weight at step $(n+1)$ (after adjustment).

The back-propagation algorithm has some drawbacks, for instance, it is sensitive to the initial values of weights as a result of its gradient descent nature. For example, Hassoun (1995) reported that the convergence may become very slow as training starts with a set of initial weights that are positioned in a flat region of the error surface, and Maier and Dandy (1998) also reported that training may start from an unfavourable position in the weight space from which the network may get stuck in a local minimum and cannot escape. Therefore, the effect of using different initial values of weights on the performance of ANN model will be investigated in the following chapters.

There are two different training modes for updating the weights, namely, the on-line mode and the batch mode. In the on-line mode, the weights are updated after each training case is presented. In contrast, in the batch mode, the network proceeds by making weight and bias changes based on entire set of input vectors. It has been suggested that the on-line mode is better than the batch mode as the sequence of training cases presented to the network can be easily randomised to avoid local minima (Zhang, 1997). Consequently, the on-line mode will be adopted for all ANN models developed in this research.

The choice of the learning rate is critical for network's training. If low value of learning rate is selected, the network learns very slowly. On the other hand, the weights and objective function diverge if higher value of learning rate is selected. Therefore, there will be no learning at all and convergence will never occur. To solve the problem, Rumelhart et al. (1986) described a process where a momentum term

(μ) is added to the weight adjustment that is proportional to the amount of the previous weight change. Once an adjustment is carried out, it is saved and used to modify all subsequent weight adjustments. This implies that the weight change of the current step should carry some momentum of the weight change from the previous step. The mathematical expressions of the modified adjustment equations are as follows

$$\Delta w_{ji}(n+1) = -\eta \frac{\partial E}{\partial w_{ji}} + \mu \Delta w_{ji}(n) \tag{3.130}$$

and

$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji}(n+1) \tag{3.131}$$

Sarle (1994a) had suggested that a momentum value of 0.9 is customarily set for both on-line and batch training modes, whilst Ripley (1993) suggested momentum values of 0.99 or 0.999 to be used for on-line training mode and a smaller value of 0.5 for batch training mode. However, Sarle (1994a) pointed out that the best momentum could be determined by trial-and-error.

The learning rate is assumed to be constant from one epoch to the next and from one weight to another in most of the algorithms described by Hertz at al. (1991). However, some researchers (e.g. Chan and Fallside, 1987; Jacobs, 1988) hold different views by proposing learning rules that vary the learning rate and provided some guidelines for learning rate update. It was found that updating learning rate can decrease the number of cycles required for training. However, it has also been found that the automatic method of updating learning rate has the risk of being trapped in local minima (Mukherjee and Deshpande, 1997).

The training process of the multi-layered perceptrons (MLPs) that are based on the back-propagation algorithms, however, requires long training time and slow to converge (Wasserman, 1989; Vitela and Reifman, 1997). This problem is attributed to the fact that these networks rely on non-linear transfer functions for learning. In more detail, if node activation is large, nodal output may tend to get stuck in the flat

spots at the extreme values of the transfer functions. The changes used to update the weights are a function of the derivative of the transfer functions. At the extreme values of the transfer functions the derivative is near zero. Consequently, very small weight changes can occur, resulting in a slow down in convergence. In order to overcome this shortcoming, Fahlman (1988) proposed adding a small, constant value to the derivative of the transfer function to prevent it from becoming zero. Falman (1988) achieved a dramatic improvement in training time by adding 0.1 to the derivative of the sigmoid transfer function. Another approach of solving the problem is the adjustment of the transfer function so that it never drops below a predefined level (Rojas, 1996).

Another known limitation of MLPs trained with the back-propagation algorithm is that the training process might be trapped in a local minimum. Weiss and Kulikowski (1991), however, suggested that local minima are not a significant problem for many applications, as they occur relatively infrequently. There are various ways proposed in the literature to avoid local minima, including increasing the learning rate, adding a momentum term, adding a small amount of random noise to the input patterns to shake the network from the line of steepest descent, adding more hidden nodes and relocating the network along the error surface, randomising the initial weights and retraining (Sietsma and Dow, 1988; Vitela and Reifman, 1997; Maier and Dandy, 2000).

Finally, despite being prolific models, feed-forward neural networks that are trained with the back propagation algorithm are often ubiquitously criticised for being *black boxes*. The knowledge acquired from these networks during training is stored in their connection weights and bias values in a complex manner that is often difficult to interpret (Touretzky and Pomerleau, 1989; Hegazy et al., 1994; Brown and Harris, 1995; Lin and Dong, 1998). Consequently, the rules governing the relationship between the input/output variables are difficult to quantify, especially for large networks that have a large number of PEs. However, work is developing in this regard to enable improved knowledge extraction form ANNs, and various approaches, such as rule extraction, contribution analysis and network inversion have been proposed.

3.9.3 Development of Artificial Neural Networks (ANNs)

There are numbers of factors that have great influences on the performance of networks. In order to improve performance, ANN models need to be developed in a systematic manner and there are a number of guidelines to address those influencing factors. Such influencing factors include the determination of model inputs, data division and pre-processing, the choice of network architecture, selection of the initial parameters that control the optimisation method, the stopping criteria and model validation (Maier and Dandy, 2000) and these factors are discussed below.

3.9.4 Model Inputs

The first process of developing an ANNs based model is selecting the model inputs. The selection of model input variables is crucial as it has the most significant impact on model performance. Presenting a large number of input variables will increase network size, resulting in a reduction in processing speed and thus the efficiency (Lachtermacher and Fuller, 1994). Various approaches have been suggested in the literature to assist with the determination of the appropriate set of input variables in the neural networks. One approach, that is usually employed in the field of geotechnical engineering, is using a fixed number of input variables in advance and assumed to be most effective input variables in relation to the model output variables (Shahin, 2003). Another approach is to train many neural networks with different combinations of input variables and compare their performance, and the network with best performance would be selected (Goh, 1994b; Najjar et al., 1996; Ural and Saka, 1998). An approach by Maier and Dandy (2000) suggested that separate networks are trained with each using only one of the variables as model input. The network that has best performance is then retrained, adding each of the remaining variables to be model input. This approach is repeated, each time with one additional input variable, until no improvement in model performance can be observed. Other approaches suggested in the literature include employing a genetic algorithm to search for the best sets of input variables (NeuralWare, 1997) and using the adaptive spline modelling of observation data (ASMOD) introduced by Kavli (1993) to develop parsimonious neural networks.

3.9.5 Division of Data

Due to their large number of parameters, ANNs can overfit the training data, particularly if they are noisy. Overfitting means that the training algorithms adjust the weights of the ANN to fit every single data value in the data set and at the same time decreases the model's capability to generalise new data sets. In other words, the model might no longer fit the general trend, but might learn the idiosyncrasies of the particular data points used for calibration, leading to 'memorisation', rather than 'generalisation'. Consequently, two separate subsets are needed: a training set, to construct the neural network model, and an independent validation set to estimate model performance in a deployed environment (Twomey and Smith, 1997; Maier and Dandy, 2000). A more efficient data division method is proposed by Stone (1974) called cross-validation in which the data are divided into three sets: training, testing and validation. The additional testing set is used to check the performance of the model at various stages of training and to determine when to stop training to avoid over-fitting.

As ANNs have difficulty extrapolating beyond the range of the data used for calibration, in order to develop the best ANN model, given the available data, all of the patterns that are contained in the data need to be included in the calibration set. For example, if the available data contain extreme data points that were excluded from the calibration data set, the model cannot be expected to perform well, as the validation data will test the model's extrapolation ability, and not its interpolation ability. If all of the patterns that are contained in the available data are contained in the calibration set, the toughest evaluation of the generalisation ability of the model is if all the patterns (and not just a subset) are contained in the validation data. In addition, if cross-validation is used as the stopping criterion, the results obtained using the testing set have to be representative of those obtained using the training set, as the testing set is used to decide when to stop training or for example which model architecture of learning rate is optimal. Consequently, the statistical properties (e.g. mean and standard deviation) of the various data subsets (e.g. training, testing and validation) need to be similar to ensure that each subset represents the same statistical population (Masters, 1993). Several studies have used ad-hoc methods to ensure that the data used for calibration and validation have the same statistical properties

(Braddock et al., 1998; Campolo et al., 1999; Tokar and Johnson, 1999; Ray and Klindworth, 2000). However, it was not until recently that systematic approaches for data division have been proposed in the literature. Bowden et al. (2002) used a genetic algorithm to minimise the difference between the means and standard deviations of the data in the training, testing, and validation sets. A recent study by Shahin et al. (2004) compared the performances of neural networks that are trained using four different data division strategies, and concluded that there is a clear direct relationship between the consistency in the statistics (i.e. mean, standard deviation, maximum and minimum) between training, testing, and validation sets and consistency in the model performance. Shahin et al. (2004) also showed that there is no clear link between the proportion of data used for training, testing and validation sets and model performance, although the best results obtained when 20% of the data used for validation and the remaining data are divided into 70% for training and 30% for testing. Shahin et al. (2004) also suggested using self-organizing map (SOM) and fuzzy clustering, which cluster data into similar groups, to decide which proportion of the available data to be used for training, testing and validation.

3.9.6 Data Pre-processing

The main objective of data pre-processing is to ensure all variables receive equal attention during the training process. Moreover, it is found that this processing usually speeds up the learning process, greatly simplifying the task of model development. Pre-processing can be in the form of data scaling, normalisation and transformation (Masters, 1993). Scaling the output data is essential, as they need to be commensurate with the limits of the transfer functions used in the output layer (e.g. between -1.0 to 1.0 for the tanh transfer function and 0.0 to 1.0 for the sigmoid transfer function). Scaling the input data is not necessary but it is almost always recommended (Masters, 1993). In some cases, the input data need to be normally distributed in order to obtain optimal results (Fortin et al., 1997). However, Burke and Ignizio (1992) stated that the probability distribution of the input data does not have to be known. Transforming the input data into some known forms (e.g. linear, log, exponential, etc.) may be helpful to improve ANN performance. Shi (2000) showed that distribution transformation of the input data to a uniform distribution improved network performance by 50%. However, empirical trials (Faraway and

Chatfield, 1998) showed that the model fits were the same, regardless of whether raw or transformed data were used.

3.9.7 Determination of Model Architecture

As is well known, determining the network architecture is one of the most important and difficult tasks in ANN model development (Maier and Dandy, 2000). It requires the selection of the optimum number of layers and the number of nodes in each of these. There is no known practical method or unified theory for determination of an optimal ANN architecture. There are always two layers representing the input and output variables, in any neural network. The intermediate layers between the input and output layer are called the hidden layers and these hidden layers are used to model relationship between input and output layers. There are some suggestions that one hidden layer is sufficient to approximate any continuous function provided that sufficient connection weights are given (Cybenko, 1989: Hornik et al., 1989). Hecht-Nielsen (1989) presented a proof that a single hidden layer of neurons, operating a sigmoidal activation function, is sufficient to model any solution surface of practical interest. To the contrary, Flood (1991) stated that there are many solution surfaces that are extremely difficult to model using a sigmoidal network containing one hidden layer. In addition, some researchers (Flood and Kartam, 1994; Sarle, 1994b; Ripley, 1996) stated that the use of more than one hidden layer provides the flexibility needed to model complex functions in many situations. Lapedes and Farber (1998) provided more practical proof that two hidden layers are sufficient, and according to Chester (1990), the first hidden layer is used to extract the local features of the input patterns while the second hidden layer is useful to extract the global features of the training patterns. However, using more than one hidden layer often slows the training process dramatically and increases the chance of getting trapped in local minima (Masters, 1993).

The number of nodes in the input and output layers are restricted by the number of the model inputs and outputs, respectively. However, there is no direct and precise way of determining the best number of nodes in each hidden layer. It is suggested that a trail-and-error procedure, which is generally used in geotechnical engineering to determine the number and connectivity of the hidden layer nodes, can be used. It has

been shown in the literature (e.g. Maren et al., 1990; Masters, 1993; Rojas, 1996) that neural networks with a large number of free parameters (connection weights) are more subject to overfitting and poor generalisation. Consequently, keeping the number of hidden nodes to a minimum, provided that satisfactory performance is achieved, is always better, as it: (a) reduces the computational time needed for training; (b) helps the network to achieve better generalisation performance; (c) avoids the problem of overfitting and (d) allows the trained network to be analysed more easily. For single hidden layer networks, there are number of approaches that can be used to obtain the best number of hidden layer nodes. One approach is to assume the number of hidden nodes to be 75% of the number of input units. (Salchenberger et al., 1992). Another approach suggests that the number of hidden nodes should be between the average and the sum of the nodes in the input and output layers (Berke and Hajela, 1991). A third approach is to fix the upper bound of the number of hidden nodes, and then work back from this bound. Hecht-Nielsen (1987) and Caudill (1988) suggested that the upper limit of the number of hidden nodes in a single layer network may be taken as $(2I + 1)$, where I is the number of inputs. The best approach found by Nawari et al. (1999) is to start with a small number of nodes and to slightly increase the number until no significant improvement in model performance is achieved. Yu (1992) showed that the error surface of a network with one hidden layer and $(I - 1)$ hidden nodes has no local minima. For networks with two hidden layers, the *geometric pyramid rule* described by Nawari et al. (1999) can be used. The notion behind this method is that the number of nodes in each layer follows a geometric progression of a pyramid shape, in which the number of nodes decreases from the input layer towards the output layer. Kudrycki (1998) found empirically that the optimum ratio of the first to second hidden layer nodes is 3:1, even for high dimensional inputs.

Another way of determining the optimal number of hidden nodes is to relate the number of hidden nodes to the number of available training samples. There are a number of rules-of-thumb that have been suggested in the literature to relate the training samples to the number of connection weights. For instance, Rogers and Dowla (1994) suggested that the number of weights should not exceed the number of training samples. Masters (1993) stated that the required minimum ratio of the number of training samples to the number of connection weights should be 2 and the

minimum ratio of the optimum training sample size to the number of connection weights should be 4. Hush and Horne (1993) suggested that this ratio should be 10, while Amari et al. (1997) suggested a value of at least 30.

More recently, a number of systematic approaches have been proposed to automatically obtain the optimal network architecture. The adaptive method of architecture determination, suggested by Ghaboussi and Sidarta (1998), is an example of an automatic method for obtaining the optimal network architecture, which suggests starting with an arbitrary, but small, number of nodes in the hidden layers. During training, and as the network approaches its capacity, new nodes are added to the hidden layers, and new connection weights are generated. Training is continued immediately after the new hidden nodes are added to allow the new connection weights to acquire the portion of the knowledge base that was not stored in the old connection weights. For this process to be achieved, some training is carried out with the new modified connection weights only, while the old connection weights are frozen. Additional cycles of training are then carried out where all the connection weights are allowed to change. The above steps are repeated and new hidden nodes are added as needed to the end of the training process, in which the appropriate network architecture is automatically determined. *Pruning* is another automatic approach to determine the optimal number of hidden nodes. One such technique proposed by Karnin (1990) starts training a network that is relatively large and later reduces the size of the network by removing the unnecessary hidden nodes. Genetic algorithms provide evolutionary alternatives to obtain optimal neural network architecture that have been used successfully in many situations (Miller et al., 1989).

3.9.8 Model Optimisation (Training)

The process of optimising the connection weights is known as ‘training’ or ‘learning’. The aim of this process is to find a global solution to what is typically a highly non-linear optimisation problem (White, 1989). The method most commonly used for finding the optimum weight combination of feed-forward neural networks is the back-propagation algorithm (Rumelhart et al., 1986), which, as mentioned previously, is based on first-order gradient descent. The main criticism of this approach is its slow asymptotic convergence rate. However, if training speed is not a major concern, there

is no reason why the back-propagation algorithm cannot be used successfully (Breiman, 1994). Consequently, the back-propagation algorithm will be employed for optimising the connection weights of the MLP models developed in chapters 6 and 7.

3.9.9 Stopping Criteria

Stopping criteria are important in the model development process as they assist in determining whether the model has been optimally trained. A number of stopping criteria can be used, including presentation of a fixed number of training epochs; a sufficiently small value of the training error; or when no or slight progress in the learning. However, the aforementioned stopping criteria may lead to the model stopping prematurely or, in some instance, over-training. As mentioned previously, Stone (1974) proposed a cross-validation technique could be used to overcome aforementioned problems. Technically, the cross-validation technique requires the data be divided into three sets: training, testing and validation. The purpose of testing set is to provide a measurement of the ability of the model to generalise, and the performance of the model using this set is checked at many stages of the training process and training is stopped when the error of the testing set starts to increase. It is considered to be the most useful tool to ensure over-fitting does not occur (Smith, 1993). The testing set is also used to determine the optimum number of hidden layer nodes and the optimum values of the internal parameters (learning rate, momentum term and initial weights). Amari et al. (1997) suggested that there are clear benefits in using cross-validation when limited data are available, as is the case for many real-life case studies. The benefits of cross-validation are discussed further by Hassoun (1995). As a result of these benefits, cross-validation will be used for the development of all MLP-based models in this research.

3.9.10 Model Validation

When the training phase of a MLP model is accomplished, the model validation phase will be carried out to ensure the model will able to generalise within the limit given by the training data in robust fashion. In validation phase, an independent validation set, which has not been used as part of the training phase, is used. If validation performance is adequate, the model is believed to be able to generalise and is

considered to be robust. Sensitivity analysis suggested by Shahin (2003) will be adopted to test the robustness and the predictive ability of MLP model. The predicted outcomes are examined against the variation of input variables and known underlying physical behaviour.

The coefficient of correlation, *r*, the root mean squared error, RMSE, and the mean absolute error, MAE, are the main criteria that are often used to evaluate the performance of MLP models. The coefficient of correlation is used to determine the relative correlation and the goodness-of-fit between the predicted and observed data and can be calculated as follows:

$$r = \frac{C_{y_j d_j}}{\sigma_{y_j} \sigma_{d_j}} \tag{3.132}$$

and

$$C_{y_j d_j} = \frac{1}{n-1} \sum_{j=1}^{n} \left(y_j - \overline{y}\right)\left(d_j - \overline{d}\right) = \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j d_j - \frac{\sum_{j=1}^{n} y_j \sum_{j=1}^{n} d_j}{n} \right) \tag{3.133}$$

$$\sigma_{y_j} = \sqrt{\frac{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2}{n-1}} \tag{3.134}$$

$$\sigma_{d_j} = \sqrt{\frac{\sum_{j=1}^{n} \left(d_j - \overline{d}\right)^2}{n-1}} \tag{3.135}$$

$$\overline{y} = \frac{\sum_{j=1}^{n} y_i}{n} \tag{3.136}$$

$$\overline{d} = \frac{\sum_{j=1}^{n} d_i}{n} \tag{3.137}$$

where:

$C_{y_j d_j}$ = the covariance between the model output and measured actual output;

y_j = the predicted output by the network;

d_j = the historical or measured actual output;

σ_{y_j} = the standard deviation of model output;

σ_{d_j} = the standard deviation of measured actual output;

$\overline{y}$ = the mean of model output;

$\overline{d}$ = the mean of measured actual output; and

n = number of data.

The RMSE is the most popular measure of error and has the advantage that large errors receive much greater attention than small errors (Hecht-Nielsen, 1990). RMSE is calculated as follows:

$$\text{RMSE} = \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(y_j - d_j \right)^2 \right\}^{1/2} \tag{3.138}$$

In contrast to RMSE, MAE eliminates the emphasis given to large errors and is calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} \left| y_j - d_j \right| \tag{3.139}$$

3.9.11 Example of ANN-based Geotechnical Model

As mentioned earlier, ANNs have been applied to solve many geotechnical engineering problems such as, for example, settlement prediction of shallow foundations on granular soils (Shahin et al., 2002). Shahin et al. (2002) have

developed MLP-based models, which outperform three of the most commonly-used traditional methods (i.e. Meyerhof, 1965; Schultze and Sherif, 1973; and Schmertmann et al., 1978) for settlement prediction of shallow foundations on granular soils. Five input parameters, including footing width, B, net applied footing load, q, average SPT blow count, N, footing length to width ratio, L/B, and footing embedment ratio, D_f/B, were used in the study, whilst the single model output was foundation settlement, S_m. In total, there were 189 individual cases in the database used for model development. These data were divided, in such a way that they are statistically consistent, into training, testing and validation sets and pre-processed (scaled between 0.0 and 1.0). Finally, an optimal model, which comprised an input layer with 5 input layer nodes, one hidden layer with two hidden layer nodes, and an output layer with a single output layer node, was found utilising trial-and-error. The structure of the authors' optimal ANN model is shown in Figure 3.16.

The connection weights and threshold levels of the neural network enable the ANN model to be translated into a relatively simple formula in which the predicted settlement can be expressed as:

$$S_m = 0.6 + \left[\frac{120.4}{1 + e^{(0.312 - 0.725 \tanh x_1 + 2.984 \tanh x_2)}} \right] \tag{3.140}$$

where:

$$x_1 = 0.1 + 10^{-3}\left[3.8B + 0.7q + 4.1N - 1.8(L/B) + 19(D_f/B)\right] \tag{3.141}$$

$$x_2 = 10^{-3}\left[0.7 - 41B - 1.6q + 75N - 52(L/B) + 740(D_f/B)\right] \tag{3.142}$$

S_m = predicted settlement (mm);

B = footing width (m);

q = net applied footing load (kPa);

N = average SPT blow count;

L/B = footing geometry; and

D_f/B = footing embedment ratio.

NOTE:
This figure is included on page 100 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.16 The structure of the optimal ANN model in Shahin et al. (2002).

A comparison carried out by Shahin et al. (2002) indicated that the MLP-based model provided more accurate settlement prediction than the traditional methods. It is also revealed that the MLP-based model has a high coefficient of correlation, *r*, and low root mean squared error, RMSE, and mean absolute error, MAE. And, more importantly, the results of the sensitivity analysis showed that the predicted settlements are in good agreement with the underlying physical behaviour of settlement based on known geotechnical knowledge, and hence, the MLP-based model is robust.

3.10 SUMMARY

Limit analysis is useful in many practical engineering areas, such as the design of mechanical structures or the analysis of soil mechanics. This method provides an estimation of the minimum load distribution that will cause a rigid, perfectly plastic solid body, which is subjected to this static load distribution, to collapse. It has demonstrated its ability to bracket the actual collapse load with sufficient accuracy from above and below. Therefore, this method will be employed in this research to estimate the bearing capacity of footings on multi-layer, variable soil profiles. The results from the limit analysis will be compared with those obtained from DFEM, not only to validate the finding from each of the numerical methods, but also to obtain a

deeper understanding of the behaviour of foundations on spatially variable soil. Two finite element limit analysis codes, UPPER and LOWER, that are provided and developed by the Geotechnical Research Group of the University of Newcastle, will be employed in all numerical analyses in this study.

It has been demonstrated that random field theory provides an appropriate framework for quantifying and estimating the spatial variability of geotechnical engineering properties of soils and rock. As mentioned previously, one of the aims of this research is to evaluate the effect of spatial variability of soils on the bearing capacity of shallow foundations. This research attempts to incorporate the local area subdivision (LAS) method with finite element implementations of lower and upper bound theorems, in addition to the DFEM.

Furthermore, it has been demonstrated that the ANNs are systems that can learn to solve complex problems from a set of known data, and generalise the acquired knowledge for predicting and forecasting applications. It is particularly useful in real-world, data-intensive applications, which deal with complex, often incomplete data. As shown previously, development of reliable and robust ANN models rely on a number of factors, which include data selection and pre-processing, the determination of suitable network architecture, selection of parameters such as stopping criteria, and finally, the validation of the model. This research will employ ANNs, in an attempt to develop a meta-model for predicting the bearing capacity of shallow foundations on multi-layered soils.