- Developing a hydraulic jump length model on horizontal rough beds
- A case for the adoption of decentralised reinforcement learning for the control of traffic flow on South African highways
- Implementation of a performance-grade bitumen specification in South Africa
- Monitoring permeability potential of hot mix asphalt via binary aggregate packing principles correlated with Bailey ratios and porosity principles
- Validating traffic models using large-scale Automatic Number Plate Recognition (ANPR) data
- Estimating elastic moduli of sandstones using two-dimensional pore space images
CONTENTS

2 Developing a hydraulic jump length model on horizontal rough beds
F Yousefi, J Mozaffari, S A M Movahed  DOI: 10.17159/2309-8775/2019/v61n3a1

7 A case for the adoption of decentralised reinforcement learning for the control of traffic flow on South African highways
T Schmidt-Dumont, J H van Vuuren  DOI: 10.17159/2309-8775/2019/v61n3a2

20 Implementation of a performance-grade bitumen specification in South Africa
S J Bredenhann, P A Myburgh, K J Jenkins, J S O’Connell, G M Rowe, J D’Angelo  DOI: 10.17159/2309-8775/2019/v61n3a3

32 Monitoring permeability potential of hot mix asphalt via binary aggregate packing principles correlated with Bailey ratios and porosity principles
E Horak, J Maina, P Myburgh, H Sebaaly  DOI: 10.17159/2309-8775/2019/v61n3a4

45 Validating traffic models using large-scale Automatic Number Plate Recognition (ANPR) data
A Robinson, C Venter  DOI: 10.17159/2309-8775/2019/v61n3a5

58 Estimating elastic moduli of sandstones using two-dimensional pore space images
T C Ekneligoda  DOI: 10.17159/2309-8775/2019/v61n3a6
Developing a hydraulic jump length model on horizontal rough beds

F Yousefi, J Mozaffari, S A M Movahed

Hydraulic jumps play an important role in the dissipation of kinetic energy downstream of hydraulic structures. The roughness of the stilling basin increases energy loss and will affect the hydraulic jump length. In this research, to estimate hydraulic jump length and consider roughness, a semi-analytical equation with unspecified coefficient was developed. Then, 244 sets of laboratory data were used to determine the coefficient of equation. The first series of data was related to a flume with a width of 20 cm and Froude number in the range of 1.1 to 5. The second series of data was obtained from a flume with a width of 50 cm and Froude number in the range of 1.02 to 9.19. USBR data was also used to increase the Froude number range as the third series of data. The results showed that the hydraulic jump length obtained from this model is a function of upstream and downstream depths, upstream Froude number, and bed roughness. It was found that the hydraulic jump length obtained from this model has an error of about 8% in comparison with observational values. In addition, according to the model, increasing roughness will reduce the length of the hydraulic jump.

INTRODUCTION

A hydraulic jump causes kinetic energy dissipation downstream of the hydraulic structures, whenever the flow regime is changed from supercritical to subcritical flow. Stilling basins are considered for energy dissipation to reduce flow energy and create conditions for the occurrence of hydraulic jump in a specific position (Jain 2001). A significant parameter to determine hydraulic jump length is the bed roughness. The results of previous studies have shown that increasing the bed roughness may cause more energy dissipation and, accordingly, may change the hydraulic jump length. Peterka (1984), in the USBR laboratory, presented a graph \( L/J_2 \) vs \( Fr_1 \) to determine hydraulic jump length based on numerous experiments in some channels. Ead and Rajaratnam (2002) investigated hydraulic jump over corrugated aluminium beds, and developed an empirical model. They found that hydraulic jump length over corrugated bed surfaces was approximately half the classical jump length. Izadjoo and Shafai-Bejestan (2007) studied hydraulic jump over four corrugated beds. Pagliara et al. (2008) determined an equation for hydraulic jump length over both uniform and non-uniform rough beds in horizontal channels. In recent years, Chanson (2009), Alikhani et al. (2010), Nasr Esfahani and Shafai-Bajestan (2012), Imran and Akib (2013), and Riazi and Jafari (2014) have studied the effect of drop and barrier height, divergence angle of stilling basins and rough and corrugated bed on the characteristics of hydraulic jump and energy dissipation using dimensional analysis. Kumar and Lodhi (2016) investigated the effect of bed roughness heights on the characteristics of a hydraulic jump. They concluded that the bed roughness height has no considerable effect on hydraulic jump characteristics. Due to the impact of roughness on hydraulic jump, the main aim of this study was to develop a new model to estimate hydraulic jump lengths on horizontal rough beds using dimensional analysis and physical model tests.

MATERIALS AND METHODS

Dimensional analysis

Prior to dimensional analysis, effective factors in physical phenomena should be determined. In the following, the parameters affecting hydraulic jump length will be determined using the momentum equation (Figure 1).

The following equation can be written:

\[ F_1 - F_2 - F_D - F_f = \rho Q(v_2 - v_1) \]  

(1)

Where: \( F_1 \) and \( F_2 \) are the hydrostatic pressure forces on the sides of the control volume between sections 1 and 2, respectively, \( F_f \) is the friction force, and \( F_D \) is the drag...
force due to the obstructions in the flow path or deformation of the channel section.

By assuming a rectangular cross-section with a width \( B \) and without obstruction in the flow path, Equation 2 can be written as follows:

\[
\frac{g y_1^2}{2} - \frac{g y_2^2}{2} - \tau_0 L_j \bar{P} = \rho B y_1 (v_2 - v_1) \tag{2}
\]

Where: \( y_1 \) and \( y_2 \) are the alternate depths, \( v_1 \) and \( v_2 \) are the velocities, \( Q \) is the flow discharge, \( \bar{P} \) is the average wetted perimeter perpendicular to flow direction, \( L_j \) is the jump length parallel to the channel bed, and \( \tau_0 \) is the average shear stress between sections 1 and 2. By replacing \( y \) by its equivalent \((a\rho v)^2\), Equation 3 can be written:

\[
\frac{\rho g y_1^2}{2} - \frac{\rho g y_2^2}{2} - \frac{k' a}{B} \left( v_1 + v_2 \right)^2 \left( \frac{v_1 + v_2}{2} \right) + \frac{\rho B y_1 (v_2 - v_1)}{2} \tag{3}
\]

In Equation 3, correction factor \( k \) is taken into account in terms of the curvature of water surface between sections 1 and 2, and for calculation of the control volume at this distance. The correction factor \( k' \) is also considered for taking into account the effect of air mixing and the reduced water volume at this distance. In addition, the coefficient “a” depends on the bed roughness. In the third term of Equation 3, \( y_1 \) and \( y_2 \) can be replaced by its equivalent \( (a\rho v)^2 \), and \( \tau_0 \) by its equivalent \( (a\rho v)^2 \). Equation 5 can be written as follows:

\[
L_j = \frac{2}{2} \frac{v_1 (v_2 - v_1)}{2} - \frac{g y_1^2 + g y_2^2}{2} - \frac{1}{2} k' a (v_1 + v_2)^2 (M + N + 1) \tag{4}
\]

Where: \( g \) is gravity acceleration, and dimensionless parameter \( a \) mainly depends on bed roughness \( k_s \). Although in Equation 4 the values of parameters \( a, k_s, k', M \) and \( N \) are unknown, the result of this equation shows the dependent factors on the jump length. Both dimensionless parameters of \( k \) and \( k' \) mainly depend on viscosity and flow velocity (or Reynolds number). In addition, the dimensionless parameters of \( M \) and \( N \) depend on the flow depth and velocity in sections 1 and 2. Regarding the components of Equation 4, these can be written as follows:

\[
f_j (v_1, v_2, v_3, g, L_j, k_s, R_e) = 0 \tag{5}
\]

In open channel flows, when the Reynolds number is greater than 2000, the effect of viscosity can be avoided (Hager and Bremen 1989). Therefore, Equation 5 can be written as follows:

\[
f_j (v_1, v_2, y_2, g, k_s, L_j) = 0 \tag{6}
\]

Equation 6 changes to an analytical equation using dimensional analysis through the Buckingham theorem. However, an unspecified coefficient is created in dimensional analysis that can be determined using experimental data and non-linear regression in 3D Table Curve software.

**Laboratory flumes**

The data used in this research was obtained from experiments performed on two hydraulic laboratory flumes in the Water Sciences and Engineering Department of Arak University. These two laboratory flumes are 20 cm and 50 cm wide, respectively, and are shown in Figure 2.

The experimental conditions are shown in Table 1 (Flume 1 and Flume 2). In addition, in order to increase the range of Froude numbers, USBR data was used (Peterka 1984). For this purpose, 89 data points were fitted using experimental data and non-linear regression in 3D Table Curve software.
sets were extracted as A, B, C, D and E flumes (Table 1).

In this study, an ultrasonic flow meter with a precision of ±0.5% was used to measure the flow discharge. In order to form a hydraulic jump, the flow in various discharges and different depths was passed through a radial gate. Then the hydraulic jump was formed by adjusting the weir installed downstream of the flume to create a suitable sequent depth. Finally, the depths before and after the hydraulic jump were measured by a point gauge. Five bed surfaces were tested in the experiments: smooth bed, $k_s = 0.16$, 0.37, 1.2 and 1.8 cm. Figure 3 shows the bed roughness used in these experiments.

For estimating the average error of the model, Equation 7 is used as follows:

$$E_i = \frac{\sum |L_o - L_p|}{L_o} \times 100$$

Where: $E_i$ is the percentage of errors, and $L_o$ and $L_p$ are respectively the observed and computed values of $L_j/y_1$.

## RESULTS AND DISCUSSION

### Dimensional analysis

Equation 6 is considered for dimensional analysis. According to the continuity equation (Equation 8), there is a definite relationship between $v_1$ and $v_2$. Therefore, one of the velocities can be eliminated from Equation 6. If $v_2$ remains, the final equation is obtained in terms of $Fr_2$. Also, if $v_1$ remains, the final equation is obtained in terms of $Fr_1$.

$$v_1 y_1 = v_2 y_2 \Rightarrow v_2 = \frac{v_1 y_1}{y_2} \text{ or } v_1 = \frac{v_2 y_2}{y_1}$$

To eliminate $v_1$ or $v_2$ in the final equation, four types of graphs are drawn (Figure 4). As shown in Figure 4, for Graph A the regression coefficient is more than 99% and the data scattering is appropriate. Also, the regression coefficients in Graph B (like USBR graph), Graph C and Graph D are 57%, 47% and 95% respectively.

According to Figure 4, the best fitness is observed for Graph A, where $L_j/y_1$ is placed versus $Fr_1$. So, $L_j/y_1$ versus $Fr_1$ can be used to determine the new model, and $v_2$ will be eliminated from Equation 6. Therefore, by the elimination of $v_2$, Equation 6 can be written as follows:

$$f_3(y_1, v_1, y_2, g, k_s, L_j) = 0$$

According to Equation 9 and the Buckingham theorem, there are four dimensionless parameters, defined as:

Figure 3 A schematic view of bed roughness used in this study

Figure 4 Data scattering of Froude number versus hydraulic jump length
\[ \Pi_1 = \frac{L}{y_1}, \Pi_2 = \frac{v_1}{y_1}, \Pi_3 = \frac{k_s}{y_2} \text{ and } \Pi_4 = \frac{v_1}{y_1} \cdot g \text{ and then } \Pi_4 = \frac{g v_1}{y_1} \]

By solving the system of equations, one can write the following relationship:

\[ (LT^{-1} y^{-m} g)^{T - m - 2} = L^{0} T^{0} \]

\[ \Rightarrow \left\{ \begin{array}{l} m + n + 1 = 0 \\ -m - 2 = 0 \end{array} \Rightarrow \left\{ \begin{array}{l} n = 1 \\ m = -2 \end{array} \right. \] (10)

By determining the unknown parameters, the fourth dimensionless parameter will be equal to:

\[ \Pi_4 = v_1 \cdot g \]

\[ f_2 \left\{ \frac{Fr_1^2}{y_1} - \frac{L_j}{y_1}, \frac{k_s}{y_2} \right\} = 0 \Rightarrow \frac{L_j}{y_1} = f_2 \frac{k_s}{y_2} \frac{Fr_1^2}{y_1} \] (11)

Since the ratio of \( y_2/y_1 \) is a function of \( Fr_1 \) this parameter can be ignored in the dimensional analysis and Equation 11 can be written as:

\[ \frac{L_j}{y_1} = F \left( \frac{k_s}{y_2}, Fr_1^2 \right) = C_j \]

(12)

\[ L_j = C_j \cdot y_1 \] (13)

The coefficient \( C_j \) was determined using 3D Table Curve software and experimental data. The three-dimensional version of this software was used to determine regression equations; therefore, by considering the \( Fr_1^2 \) in the x-axis, \( k_s/y_2 \) in the y-axis and the measured values of \( C_j \) in the z-axis values, the best fitted curve equation is determined. Figure 5 shows the diagram for data fitting in the Table Curve software.

The selected equation for fitted surface is as follows:

\[ C_j = 27.65 + 0.45 Fr_1^2 - 48.61 \frac{k_s}{y_2} \] (14)

As can be seen, the accuracy of the determined coefficient provided by the software is more than 89% and the standard error is less than 12%.

By replacing Equation 14 by the coefficient \( C_j \) in Equation 13, the hydraulic jump length can be achieved. Therefore, the final equation for determining the hydraulic jump length is as follows:

\[ L_j = y_1 \left( 27.65 + 0.45 Fr_1^2 - 48.61 \frac{k_s}{y_2} \right) \] (15)

Equation 15 shows that the hydraulic jump length is a function of the roughness and hydraulic conditions before and after the jump. A total of 80 data sets were used for verification of the model. To do this, the observed hydraulic jump length was considered in the horizontal axis, and the jump length computed by Equation 15 was transferred to the vertical axis. As can be seen in Figure 6, the results are close to the identity line (\( y = x \)). Also, according to Equation 7, the percentage error obtained was equal to 8%.

**CONCLUSION**

Roughness is one of the most effective factors in hydraulic jump length. With increasing roughness in the stilling basin, the flow resistance and energy dissipation increase. Therefore, the jump length will change. The main purpose of this research was to develop a model for hydraulic jump length considering the roughness effect. First of all, the momentum equation was used to determine the effective parameters on hydraulic jump length. Then, by measuring the dimensional analysis and Buckingham theorem, the hydraulic jump length model was determined with a coefficient. This coefficient is a function of bed roughness, secondary depth, and Froude number. To determine this coefficient, Table Curve software and experimental data were used. The experimental data included two data series obtained in this study and a series of data from USBR experiments. USBR data was used to increase the range of the Froude number. The best equation was achieved in terms of simplicity of equation, with high regression and low standard error using Table Curve 3-D software. After that, in order to estimate the percentage error of the model, the computed values of jump length were compared with the observed values. Comparison of the computed values of jump length by the new model with the observed values, shows that the model percentage error is very low. Therefore, this model can be used to design a stilling basin, considering the bed roughness.

**REFERENCES**


A case for the adoption of decentralised reinforcement learning for the control of traffic flow on South African highways

T Schmidt-Dumont, J H van Vuuren

As an alternative to capacity expansion, various dynamic highway traffic control measures have been introduced. Ramp metering and variable speed limits are often considered to be effective dynamic highway control measures. Typically, these control measures have been employed in conjunction with either optimal control methods or online feedback control. One shortcoming of feedback control is that it provides no guarantee of optimality with respect to the chosen metering rate or speed limit. Optimal control approaches, on the other hand, are limited in respect of their applicability to large traffic networks due to their significant computational expense. Reinforcement learning is an alternative solution approach, in which an agent learns a near-optimal control strategy in an online manner, with a smaller computational overhead than those of optimal control approaches. In this paper an empirical case is made for the adoption of a decentralised reinforcement learning approach towards solving the control problems posed by both ramp metering and variable speed limits simultaneously, and in an online manner. The effectiveness of this approach is evaluated in the context of a microscopic traffic simulation model of a section of the N1 national highway outbound from Cape Town in South Africa’s Western Cape Province.

INTRODUCTION

Highways were originally built with the aim of providing virtually unlimited mobility to road users. The ongoing dramatic expansion of car ownership and travel demand has, however, led to the situation where, today, traffic congestion is a significant problem in major metropolitan areas all over the world (Schrank et al 2012). The reason for the severe traffic congestion experienced around the world is over-utilisation of the existing road infrastructure which potentially leads to dense, stop-and-go traffic. Although traffic congestion is typically associated with well-developed countries such as the United States, China or Germany, it is also a major problem in South Africa. According to the TomTom Traffic Index (TomTom 2017), a congestion ranking based on GPS data collected from individual vehicles, Cape Town is the 48th most congested city in the world, and the most congested city in Africa. In order to place these statistics into perspective, Cape Town has the same congestion ranking as New York City according to the TomTom Traffic Index published at the end of 2016. The morning and afternoon peak congestion in Cape Town furthermore exceeds that experienced by commuters in New York City.

Traffic congestion levels in Cape Town have increased steadily since 2011, with a significant increase in congestion levels from 30% in 2015 to 35% in 2016 (TomTom 2017). These percentages imply that a journey would have taken, on average, 35% longer in 2016 due to congestion than it would have taken if free-flowing traffic conditions had prevailed. During the morning and afternoon peaks, the levels of traffic congestion are naturally higher than these average values suggest. Travellers experience a 75% increase in travel time during the morning peak, while commuters experience a 67% increase in travel time during the afternoon peak. The result of these levels of traffic congestion is that the average Capetonian spends an additional 42 minutes stuck in traffic per day, which accumulate to approximately 163 hours stuck in traffic per year (TomTom 2017).

In Figure 1 it is shown that congestion levels in Johannesburg temporarily decreased from 2009 to 2012. This decrease may be
attributed to capacity expansion as a result of the Gauteng Freeway Improvement Project (SANRAL 2009). The subsequent rise in congestion levels during the period 2012–2016, visible in the figure, may be attributed to the so-called theory of induced traffic demand, in which it is suggested that increases in highway capacity will induce additional traffic demand, thus not permanently alleviating congestion as envisioned (Noland 2001).

The alternative to capacity expansion aimed at improving traffic flow on highways is more effective control of the existing infrastructure. Ramp metering (RM) is a means of improving highway traffic flow through effective regulation of the flow of vehicles that enter a highway traffic flow from an on-ramp. In this way, the mainline throughput may be increased by an avoidance of capacity loss and blockage of on-ramps as a result of congestion (Papageorgiou & Kotsiolas 2000). Variable speed limits (VSLs) were initially employed mainly with the aim of improving traffic safety on highways due to the resulting homogenisation of traffic flow (Hegyi et al 2005). In more recent developments, however, VSLs have been employed as a traffic flow optimisation technique with the aim of improving traffic flow along highways. This improvement may take one of two forms, either maintaining stable traffic flow by slightly reducing the speed limit in order to reduce the differences in speed between vehicles and reduce the following distance, resulting in improved traffic flow (Hegyi et al 2005), or by decreasing the speed limit to such an extent that an artificial bottleneck is created, inducing controlled congestion, but maintaining free-flow traffic at the true bottleneck location (Carlson et al 2010). RM and VSLs are considered to be effective highway traffic control measures (Papageorgiou & Kotsiolas 2000). An empirical case is made in this paper for their adoption within a South African context. Traditionally, classical feedback control theory has been employed in the design of controllers for implementing RM and VSLs (Carlson et al 2014). One drawback of the classical feedback control approach is that it provides no guarantee of optimal control. Furthermore, feedback controllers are purely reactive, which may result in delayed response.

Reinforcement learning (RL) provides a promising framework addressing these issues. The objective in this paper is to compare, for the first time, the relative effectiveness of state-of-the-art feedback controllers from the literature with that of employing a decentralised RL approach towards solving the RM and VSL problems simultaneously in the context of a real-world scenario. Furthermore, this paper contains, to our best knowledge, the first application of multi-agent reinforcement learning (MARL) approaches at several consecutive on-ramps contained within the same simulation model.

LITERATURE REVIEW

In this section a review of RM and VSL controllers from literature is provided, followed by a brief introduction to RL.

Highway traffic control measures

Wattleworth (1967) introduced the first RM strategies, which were based on historical traffic demand at on-ramps, setting specific metering rates for certain time intervals in order to control the inflow of traffic onto the highway. In search of a more adaptive RM strategy, Papageorgiou et al (1991) introduced the well-known Asservissement Lineaire d’entrée Autoroutiere (ALINEA) control mechanism, which is based on online feedback control theory. An extension of the ALINEA control strategy, called PI-ALINEA, was later introduced by Wang et al (2014) such that bottlenecks occurring further downstream than the immediate lane merge may also be taken into account. Alternative existing RM solutions include a model predictive control (MPC) approach proposed by Hegyi et al (2005) and an implementation of a hierarchical control approach by Papamichail et al (2010). Early RM approaches, however, often led to the formation of long queues of vehicles building up on the on-ramp, which may cause congestion in the arterial network. This issue was addressed by Smaragdis and Papageorgiou (2003) who designed an extension to be implemented in conjunction with a feedback controller (such as ALINEA) which, in cases of severe congestion, maintains a maximum on-ramp queue length set to some pre-specified value. A second metering rate is calculated, ensuring the maximum allowable queue length is not exceeded, and the least restrictive metering rate is then applied (Smaragdis & Papageorgiou 2003).

An early attempt at employing RL to solve the RM problem with the aim of learning optimal control policies in an online manner is due to Davarjennad et al (2011). They employed the well-known Q-Learning RL algorithm (Watkins & Dayan 1992) in order to learn optimal metering rates within the context of a macroscopic traffic simulation model developed in the well-known METANET traffic modelling software, while simultaneously considering the build-up of on-ramp queues. Rezaee et al (2013) demonstrated the first application of RL for solving the RM problem in the context of a microscopic traffic simulation model, in which a portion of Highway 401 in Toronto, Canada was considered.

Smulders (1990) demonstrated one of the first applications of VSLs as an optimisation technique. In his formulation of the VSL control problem as an optimal control problem, which was based on a macroscopic traffic simulation model, the aim was to determine speed limits such that the expected time until traffic congestion occurs is maximised. Alessandri et al (1998, 1999) later on extended and refined this original optimal control approach. Carlson et al (2011) subsequently proposed an online feedback controller. The controller receives real-time traffic flow and density measurements as input, which are subsequently used to calculate appropriate speed limit.
throughput may be achieved for various scenarios of traffic demand. A simpler version of such an online feedback controller was later introduced by Müller et al. (2015).

Zhu and Ukkusuri (2014), as well as Walraven et al. (2016), have shown that the VSL problem may be solved using RL techniques. Zhu and Ukkusuri (2014) demonstrated an application of the R-Markov Average Reward Technique (R-MART) RL algorithm for solving the VSL control problem within the context of a macroscopic link-based dynamic network loading model. Walraven et al. (2016), on the other hand, employed Q-Learning in conjunction with a neural network for function. In both the studies mentioned above, the VSL problem was addressed within a macroscopic traffic modelling paradigm. This paradigm may, however, be limiting as it is often difficult to replicate some of the important, realistic characteristics of traffic flow, including shockwave propagation, or the spill-back effect which may occur due to heavy congestion (Zhu & Ukkusuri 2014).

Carlson et al. (2014) proposed an integrated feedback controller for simultaneously performing both RM and enforcing VSLs. This controller comprises two individual feedback controllers. The RM controller operates according to the PI-ALINEA feedback controller with the addition of the queue limitation as defined by Smaragdis and Papageorgiou (2003). RM is then applied by itself until the on-ramp queue limit is reached, at which point a VSL controller, such as that of Carlson et al. (2011) or Müller et al. (2015), is employed in order to provide supplementary highway traffic flow control.

In the RL implementations for RM and VSLs mentioned above, the RL approaches were typically able to outperform the corresponding feedback controllers (Rezaee et al. 2013; Walraven et al. 2016). RL has, however, not been employed for solving the RM and VSL control problems simultaneously within the context of a real-world case study. It is envisioned that a MARL approach for simultaneously employing RM and VSLs may lead to further improvements in the travel times experienced by motorists. RL followed by its implementations in this paper, working towards the MARL approach, is outlined in the following sections.

Reinforcement learning
RL is the concept of learning an optimal control policy by trial and error (Sutton & Barto 1998). A learning agent receives information about the current state of the environment in which it operates at each time step. This state of the environment is typically defined by one or a number of descriptive state variables. Based on this state information, the agent performs an action that subsequently transforms the environment in which it finds itself into a new state. An agent’s behaviour is defined by its policy, which is the mapping according to which the agent chooses its action based on the current state information. The agent receives feedback in the form of a scalar reward so as to provide it with an indication of the quality of the action chosen. This reward is typically determined based on the new state of the environment, according to a specific reward function. The aim of an RL agent is to learn a policy according to which the accumulated reward that it receives over time is maximised (i.e. finding a policy that results in the agent choosing the best action in each state with respect to the long-term cumulative reward achieved) (Szepesvari 2010).

**REINFORCEMENT LEARNING IMPLEMENTATIONS**

**Reinforcement learning for RM**
Davarjenad et al. (2011) and Rezaee et al. (2013) demonstrated that RL techniques may be employed for solving the RM problem. RM is typically enforced by a traffic signal located at an on-ramp, employing a one-vehicle-per-green-phase metering protocol (Hegyi et al. 2005). The traffic signal is thus given a fixed green phase time of three seconds, allowing a single vehicle to pass during every green phase, while the RL agent controls the red phase times, thereby regulating the flow of vehicles allowed to join the highway traffic flow. An RL agent is thus required to control the traffic light at each on-ramp where RM is enforced.

**The state space**
As may be seen in Figure 2, the state space of the RM agent comprises three variables. The first of these is the density \( \rho_{ds} \), measured directly downstream of the on-ramp. It is expected that this variable will provide the agent with explicit feedback in respect of the quality of the previous action, because the merge of the on-ramp and highway traffic flows is expected to be the source of congestion. Therefore, the downstream density is expected to be the earliest indicator of impending congestion. The density \( \rho_{us} \), measured upstream of the on-ramp is the second state variable. The upstream density is included in the state space, because it provides information on the propagation of congestion backwards along the highway.

---

**Figure 2** The state space for the RM agents

\[
\begin{array}{cc}
\rho_{us} & \rho_{ds} \\
\hline
\hline
\omega & \\
\end{array}
\]
The on-ramp queue length \( w \) is the third state variable, which was included to provide the learning agent with an indication of the on-ramp demand.

### The action space

Based on the current state of traffic flow on the highway, the learning agent may select a suitable action. In this study, direct action selection (i.e., directly choosing a red phase duration from a set of pre-specified red times) is applied. As in the implementation of Rezaee et al. (2013), the agent may choose an action \( a \in \{0, 2, 3, 4, 6, 8, 10, 13\} \), where each action represents a corresponding red phase duration in seconds. Assuming a fixed green phase duration of three seconds, the key focus area where a VSL is applied. It is expected that the most immediate response to an action will be reflected on this section of a highway. Therefore, this variable should provide the agent with an indication of the effectiveness of the chosen action. Finally, the third state variable is the density measured on the highway section further upstream from that comprising the area considered for the second state variable, denoted by \( \rho_{app} \). This variable is chosen primarily to provide a predictive component in terms of highway demand. Furthermore, this variable is expected to provide the agent with an indication of the severity of the congestion in cases where it has spilled back beyond the application area of \( \rho_{app} \).

### The reward function

When designing a traffic control system, the typical objective is to minimise the total travel time spent in the system by all transportation users. From the fundamental theory of traffic flow it follows that maximum throughput occurs at the critical density of a specific highway section (Papageorgiou & Kotsialos 2000). Therefore, an RM agent typically aims to control the density on the highway. This is the case in ALINEA, the most celebrated RM technique (Rezaee et al. 2013). The reward function employed in this paper was inspired by the ALINEA control law. According to the ALINEA control strategy, the metering rate employed at an on-ramp is adjusted based on the difference between a desired downstream density and the measured downstream density (Papageorgiou et al. 1991). Furthermore, an additional punishment is included in the reward function to refrain the agent from enforcing metering rates which lead to the build-up of long on-ramp queues. The reward of the RM RL agents is therefore calculated as:

\[
    r(t) = \begin{cases} 
        -(\rho^* - \rho_{ds}(t))^2 & \text{if } w < w^* \\
        -(\rho^* - \rho_{ds}(t))^2 - 100000 & \text{if } w \geq w^*
    \end{cases}
\]

where \( \rho^* \) denotes the desired downstream density that the RL agent aims to achieve, while \( \rho_{ds}(t) \) denotes the actual downstream density measured during the last control interval, \( t \), \( w \) denotes the current measured on-ramp queue length, and \( w^* \) denotes the maximum allowable on-ramp queue length. In order to provide amplified negative feedback to the agent for actions that result in large deviations from the target density, this difference is squared. Both the Q-Learning (Watkins & Dayan 1992) and \( kNN-TD \) learning (Martin et al. 2011) algorithms are implemented for RM in this paper.

### Reinforcement learning for VSLs

Zhu and Ukkusuri (2014) and Walraven et al. (2016) have demonstrated formulations of the VSL problem as RL problems, and subsequently solved these using RL algorithms. We apply VSLs in the vicinity of each of the on-ramps. The VSLs determined by the RL agent are typically enforced by displaying the current speed limit on a roadside variable message sign.

### The state space

Similarly to the RM implementations, the state space for the VSL implementations comprises three variables, as shown in Figure 3. The first of these state variables is, as for the RM implementations, the density \( \rho_{ds} \) directly downstream of the on-ramp. This variable is again chosen so as to provide the VSL agent with information on the state of traffic flow at the bottleneck location. The density directly upstream of the bottleneck location, denoted by \( \rho_{app} \), is the second state variable. This key focus area where a VSL is applied. It is expected that the most immediate response to an action will be reflected on this section of a highway. Therefore, this variable should provide the agent with an indication of the effectiveness of the chosen action. Finally, the third state variable is the density measured on the highway section further upstream from that comprising the area considered for the second state variable, denoted by \( \rho_{us} \). This variable is chosen primarily to provide a predictive component in terms of highway demand. Furthermore, this variable is expected to provide the agent with an indication of the severity of the congestion in cases where it has spilled back beyond the application area of \( \rho_{app} \).

### The action space

Similarly to the RM implementation, direct action selection is employed in the VSL implementation. We applied the VSL:

\[
    VSL_{app} = 90 + 10a
\]

![Figure 3 The state space for the VSL agents](image-url)
where \( a \in \{0, 1, 2, 3\} \) applies. As a result, minimum and maximum variable speed limits of 90 km/h and 120 km/h respectively may be applied at the application area. The lower limit, as well as the increment of 10 km/h, was empirically determined to achieve the best performance. In order to reduce the difference in speed limit from 120 km/h to VSL\(_{\text{app}}\) the speed limit at the upstream section is adjusted according to:

\[
\text{VSL}_{\text{us}} = \max[(\text{VSL}_{\text{app}} + \delta), 120]
\]

It is envisioned that this more gradual reduction in the speed limit will reduce the probability of shock-waves which may be a result of sharp, sudden reductions in the speed limit propagating backwards along the highway.

The reward function

The VSL agent also aims to minimise the total time spent in the system by all transportation users, by maximising the system throughput. Therefore, the VSL agent is rewarded according to the flow rate out of the bottleneck location, measured in veh/h. Similarly to the RM implementations, Q-Learning (Watkins & Dayan 1992) and ANN-TD learning (Martin et al 2011) are implemented for VSLs.

### Multi-agent reinforcement learning for combined RM and VSLs

We also consider three approaches towards simultaneously solving the RM and VSL problems by means of multi-agent reinforcement learning (MARL) (Busoniu et al 2008). Employing independent learners (El Tantawy et al 2013) is the first and the simplest of these approaches, where both the RM and VSL agents learn independently, without any form of communication between them, as they both aim simply to maximise their own, local rewards. In the second approach, henceforth referred to as the hierarchical MARL approach, a hierarchy of learning agents is established. Action selection is performed according to the order in this hierarchy (i.e. the agent assigned the highest rank may choose first, followed by the second-highest ranked agent, and so forth) (Busoniu et al 2008). Once the highest ranked agent has chosen its action, this action is communicated to the second-highest ranked agent. The second-highest ranked agent can then take this action into account and select its own action accordingly. As a result of this communication, the state-action space of the second agent grows by a factor equal to the number of actions available to the first agent. These rankings may be determined empirically, or in the order of performance of the individual agents. The maximax MARL approach, which is the third and most sophisticated MARL approach, is based on the principle of locality of interaction among agents (Nair et al 2005). According to this principle, an estimate of the utility of a local neighbourhood maps the effect of an agent’s actions to the global value function (only the neighbouring agents are considered) (El Tantawy et al 2013). The implementation works as follows:

1. Each agent \( i \) chooses an action which is subsequently communicated to its neighbouring agent \( j \).
2. Each agent \( i \) finds an action \( a'_{j(i)} \) that maximises the joint gain in rewards.
3. This joint gain is calculated for each agent \( i \) as if it were the only agent allowed to change its action, while its neighbour’s action remains unchanged.
4. The agent able to achieve the largest joint gain changes its action, while the action of the neighbour remains unchanged.
5. The process is repeated during each learning iteration.

Due to this two-way communication, the state-action space of each agent increases by a factor equal to the number of actions available to its neighbour in the maximax MARL approach. The following section is devoted to a discussion of the highway traffic simulation model in which these control measures and algorithms were implemented.

### THE MICROSCOPIC TRAFFIC SIMULATION MODEL

In this section, a description of the microscopic traffic simulation model, which is used as the algorithmic test bed, is provided, illustrating the modelling tools employed, as well as the model calibration based on input data, and the techniques for analysis of the output data.

#### Modelling tools employed and case study area

A simulation model was developed as algorithmic test bed within the AnyLogic 7.3.5 University Edition (AnyLogic 2017) software suite, making specific use of its built-in Road Traffic and Process Modelling Libraries. The road traffic library allows for microscopic traffic modelling, where each vehicle is simulated individually.

The highway section modelled is a stretch of the N1 national highway outbound from Cape Town in South Africa’s Western Cape Province, from just before the R300 off-ramp (denoted by \( O_1 \)) up to a section after the on-ramp at the Okavango Road interchange (denoted by \( D_2 \)), as shown in Figure 4. Five on- and off-ramps fall within this study area, namely the off-ramp at the R300 interchange (denoted by \( D_3 \)), the on-ramp at the Brackenfell Boulevard interchange (denoted by \( O_3 \)), the off-ramp at the Okavango Road interchange (denoted by \( D_4 \)), and the on-ramp at the Okavango Road interchange (denoted by \( O_4 \)). This stretch of highway experiences significant congestion problems, especially during the afternoon peak, when large traffic

![Figure 4](image-url)
volumes enter the N1 from the R300 and leave the N1 at the Okavango Road off-ramp.

**Model input data**

Model input data was required for calibration and validation of the simulation model. This data was obtained from two major sources. The primary sources were Wavetronix® (Wavetronix 2017) smart sensor devices installed at various locations along the major highways throughout the study area, as may be seen in Figure 4. Such a sensor employs two radar beams in order to detect individual vehicles as they pass the sensor. Vehicles are classified by the sensor into three major classes, based on their respective lengths (Committee of Transport Officials 2013). These classes are (1) passenger vehicles, (2) light delivery vehicles and (3) trucks. The secondary sources of vehicle demand data were video recordings from a CCTV camera installed at a major intersection. The CCTV footage was used to estimate on- and off-ramp flows at intersections in cases where these flows could not be derived from the sensor data. These flows were estimated by human interpretation with data tagging. The sensor data was aggregated into 10-minute intervals, providing numeric values for vehicle class counts, as well as average vehicle speed data for each 10-minute interval. The sensor data was received for the entire month of March 2017, while video recordings of the afternoon peak from 15:30 to 18:30 were received for the first three Fridays of March 2017.

**General specifications of the simulation model**

In the simulation model, vehicle arrivals follow a Poisson distribution with an input mean equal to a predetermined desired traffic volume (measured in veh/h), based on the real-world volumes. It was found that traffic demand could be replicated accurately when these desired traffic volumes are adjusted in the simulation model in 30-minute intervals. As part of the calibration of the simulation model, the vehicle properties were adjusted, as these parameters have an influence on vehicle behaviour which, in turn, affects the vehicle throughput. Passenger vehicle lengths were fixed at 5 m, while light delivery vehicles were taken as 10 m, and trucks were assumed to be 15 m in length. These vehicle lengths are in line with the data collection standards set out in the Committee of Transport Officials (2013). The initial speeds for passenger vehicles entering the network at O1 and O2 were set to 100 km/h, while the corresponding initial speeds at O3 and O4 were set to 60 km/h. Similarly, light delivery vehicles entering the network at O1 or O2 were assumed to have an initial speed of 100 km/h, while light delivery vehicles entering the network at O3 or O4 were given an initial speed of 60 km/h. Finally, the initial speed of trucks entering the network at O1 or O2 was taken as 80 km/h, with trucks entering the network at a speed of 60 km/h at O3 and O4.

In order to account for different driving styles and variation in driver aggressiveness, the preferred speeds of passenger vehicles were distributed uniformly between 110 km/h and 130 km/h, while the preferred speeds of light delivery vehicles were uniformly distributed between 90 km/h and 110 km/h. Finally, the preferred speeds of trucks were distributed uniformly between 70 km/h and 90 km/h. The maximum acceleration and deceleration values for passenger vehicles were taken as 2.7 m/s² and –4.4 m/s², respectively. For light delivery vehicles these values were set to 1.5 m/s² and –3.1 m/s², respectively, while for trucks these values were set at 1.5 m/s² and –2.8 m/s², respectively. Throughout the process of adjusting these values empirically, care was taken to stay within the reasonable bounds of 1.5 m/s² to 4 m/s² for the maximum acceleration and –1 m/s² to –6 m/s² for the maximum deceleration, respectively, as suggested by Amirjamshidi and Roorda (2017) in their multi-objective approach to traffic microsimulation model calibration.

**Model validation**

The simulation model was executed for validation purposes over a period of three hours and forty minutes, so as to include a 40-minute warm-up period, before starting to record vehicle counts over the subsequent three hours. The length of the warm-up period was determined according to the method outlined by Law and Kelton (2000), ensuring that a stable number of vehicles is present in the simulation model before data recording commences. This process was replicated thirty times with different seed values. The measured outputs at the sensor locations were then compared with the real-world values of the corresponding time period from 15:30 to 18:30. The results of this comparison are shown in Figure 5.

As may be seen in Figure 5, the simulated outputs, indicated in red, resemble the real-world measurements, indicated in blue. Note that neither the measured data nor the simulated output resembles the classical fundamental diagram. This may be due to the fact that only data from a specific, congestion-plagued period of time is shown. Based on the central limit theorem, one may assume that this data is normally distributed due to the large number of recorded values. Hypothesis tests were performed in order to ensure that the means of the simulated output and the real-world values do not differ statistically at a 5% level of significance.

Furthermore, the average output results of these thirty replications were compared against the real-world measurements for all sensor and estimated locations from the
first three Fridays of March 2017, and the absolute errors were recorded, as shown in Tables 1 and 2. As may be seen in the tables, the errors in respect of the flow of passenger vehicles, abbreviated in the tables as PV, never exceeded 2% during the simulated three-hour period. In terms of the light delivery vehicles, abbreviated in the tables as LDV, the maximum error during the three simulation hours rose to 4.90%. The reason for this is that the number of light delivery vehicles travelling through the system was significantly smaller than that of passenger vehicles, resulting in the phenomenon where even a small deviation in terms of the number of vehicles is reflected as a relatively large error when expressed as a percentage.

### Table 1
Validation of simulated traffic flow at Sensors 1 and 2 as well as the Brackenfell Boulevard on-ramp

<table>
<thead>
<tr>
<th>Time period</th>
<th>Sensor 1</th>
<th>Sensor 2</th>
<th>Brackenfell Boulevard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PV</td>
<td>LDV</td>
<td>T</td>
</tr>
<tr>
<td>15:30–15:40</td>
<td>2.01%</td>
<td>8.86%</td>
<td>1.37%</td>
</tr>
<tr>
<td>15:40–15:50</td>
<td>3.48%</td>
<td>2.83%</td>
<td>16.79%</td>
</tr>
<tr>
<td>15:50–16:00</td>
<td>1.57%</td>
<td>1.30%</td>
<td>0.95%</td>
</tr>
<tr>
<td>16:00–16:10</td>
<td>0.37%</td>
<td>3.01%</td>
<td>5.47%</td>
</tr>
<tr>
<td>16:10–16:20</td>
<td>0.69%</td>
<td>0.01%</td>
<td>3.78%</td>
</tr>
<tr>
<td>16:20–16:30</td>
<td>1.27%</td>
<td>1.25%</td>
<td>4.15%</td>
</tr>
<tr>
<td>16:30–16:40</td>
<td>0.36%</td>
<td>2.03%</td>
<td>3.26%</td>
</tr>
<tr>
<td>16:40–16:50</td>
<td>0.21%</td>
<td>0.78%</td>
<td>0.12%</td>
</tr>
<tr>
<td>16:50–17:00</td>
<td>0.14%</td>
<td>0.13%</td>
<td>0.68%</td>
</tr>
<tr>
<td>17:00–17:10</td>
<td>0.49%</td>
<td>0.80%</td>
<td>0.82%</td>
</tr>
<tr>
<td>17:10–17:20</td>
<td>0.60%</td>
<td>0.44%</td>
<td>0.67%</td>
</tr>
<tr>
<td>17:20–17:30</td>
<td>1.28%</td>
<td>0.23%</td>
<td>1.24%</td>
</tr>
<tr>
<td>17:30–17:40</td>
<td>1.31%</td>
<td>0.25%</td>
<td>2.76%</td>
</tr>
<tr>
<td>17:40–17:50</td>
<td>0.80%</td>
<td>1.01%</td>
<td>2.84%</td>
</tr>
<tr>
<td>17:50–18:00</td>
<td>0.78%</td>
<td>0.40%</td>
<td>1.10%</td>
</tr>
<tr>
<td>18:00–18:10</td>
<td>0.65%</td>
<td>0.04%</td>
<td>1.13%</td>
</tr>
<tr>
<td>18:10–18:20</td>
<td>0.52%</td>
<td>0.31%</td>
<td>0.52%</td>
</tr>
<tr>
<td>18:20–18:30</td>
<td>0.33%</td>
<td>0.66%</td>
<td>2.05%</td>
</tr>
<tr>
<td>Total</td>
<td>0.24%</td>
<td></td>
<td>1.23%</td>
</tr>
</tbody>
</table>

### Table 2
Validation of simulated traffic flow at the Okavango Road off-ramp, on the N1 after the Okavango Road off-ramp and at Sensor 3

<table>
<thead>
<tr>
<th>Time period</th>
<th>Okavango Road off-ramp</th>
<th>N1 after Okavango Road off-ramp</th>
<th>Sensor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PV</td>
<td>LDV</td>
<td>T</td>
</tr>
<tr>
<td>15:30–15:40</td>
<td>11.78%</td>
<td>40.41%</td>
<td>30.67%</td>
</tr>
<tr>
<td>15:40–15:50</td>
<td>14.50%</td>
<td>41.78%</td>
<td>32.33%</td>
</tr>
<tr>
<td>15:50–16:00</td>
<td>10.29%</td>
<td>36.98%</td>
<td>24.76%</td>
</tr>
<tr>
<td>16:00–16:10</td>
<td>7.16%</td>
<td>31.48%</td>
<td>17.45%</td>
</tr>
<tr>
<td>16:10–16:20</td>
<td>5.42%</td>
<td>24.38%</td>
<td>11.33%</td>
</tr>
<tr>
<td>16:20–16:30</td>
<td>5.01%</td>
<td>15.04%</td>
<td>8.84%</td>
</tr>
<tr>
<td>16:30–16:40</td>
<td>5.38%</td>
<td>17.14%</td>
<td>11.31%</td>
</tr>
<tr>
<td>16:40–16:50</td>
<td>5.09%</td>
<td>18.15%</td>
<td>13.33%</td>
</tr>
<tr>
<td>16:50–17:00</td>
<td>3.62%</td>
<td>16.40%</td>
<td>10.67%</td>
</tr>
<tr>
<td>17:00–17:10</td>
<td>3.26%</td>
<td>13.27%</td>
<td>13.08%</td>
</tr>
<tr>
<td>17:10–17:20</td>
<td>2.05%</td>
<td>11.39%</td>
<td>10.68%</td>
</tr>
<tr>
<td>17:20–17:30</td>
<td>1.15%</td>
<td>9.37%</td>
<td>4.86%</td>
</tr>
<tr>
<td>17:30–17:40</td>
<td>1.07%</td>
<td>7.70%</td>
<td>5.36%</td>
</tr>
<tr>
<td>17:40–17:50</td>
<td>0.68%</td>
<td>6.81%</td>
<td>6.90%</td>
</tr>
<tr>
<td>17:50–18:00</td>
<td>0.18%</td>
<td>7.78%</td>
<td>8.91%</td>
</tr>
<tr>
<td>18:00–18:10</td>
<td>0.05%</td>
<td>5.15%</td>
<td>9.55%</td>
</tr>
<tr>
<td>18:10–18:20</td>
<td>0.24%</td>
<td>3.59%</td>
<td>7.58%</td>
</tr>
<tr>
<td>18:20–18:30</td>
<td>0.20%</td>
<td>3.25%</td>
<td>2.86%</td>
</tr>
<tr>
<td>Total</td>
<td>0.44%</td>
<td></td>
<td>0.71%</td>
</tr>
</tbody>
</table>
Finally, the largest error during the simulated three-hour period in terms of trucks travelling through the system, abbreviated in the tables as T, was 2.86%. As in the case of light delivery vehicles, however, relatively few trucks travelled through the system, and as a result a small error in terms of number is reflected as a relatively large error when expressed as a percentage. Due to the fact that the total error in respect of the number of vehicles that passed any of the six counting stations never exceeded 2%, the simulation model was deemed to be a sufficiently accurate representation of the underlying real-world system.

**Model output data**
The relative performances of the RL algorithms are measured according to the following performance measures (all measured in vehicle hours):
1. The total time spent in the system by all vehicles (TTS)
2. The total time spent in the system by vehicles entering the system from the N1 (TTSN1)
3. The total time spent in the system by vehicles entering the system from the R300 (TTSR300)
4. The total time spent in the system by vehicles entering the system from the Brackenfell Boulevard on-ramp (TTSBB), and
5. The total time spent in the system by vehicles entering the system from the Okavango Road on-ramp (TTSO).

The reason for breaking the TTS performance measure indicator down into the four further performance measures is that increases in the travel times of vehicles joining the highway from on-ramps at which RM is applied may not be captured sufficiently if only a single TTS performance measure were to be adopted. Now that the data inputs and outputs have been outlined, the focus of the discussion shifts to the calibration and validation of the simulation model.

**Numerical Experimentation**
The process followed throughout the numerical results evaluation is as follows. An Analysis of Variance (ANOVA) (Montgomery & Runger 2011) is performed in order to ascertain whether the simulation outputs from the different implementations differ statistically. Thereafter, Levene’s Test (Schult 1985) is performed in order to determine whether the variances of the output data sets are homogenous or not. If these variances are in fact homogenous, the Fischer LSD (Williams & Abdi 2010) *post hoc* test is employed in order to determine between which pairs of algorithmic output the differences occur. If, however, these variances are not homogenous, the Games-Howell (Games & Howell 1976) *post hoc* test is performed for this purpose.

**Ramp metering**
RM may be applied at all three on-ramps of the stretch of the N1 highway in Figure 4, namely the R300 on-ramp at O2, the Brackenfell Boulevard on-ramp at O3 and the Okavango Road on-ramp at O4, as may be seen in Figure 6. As a benchmark for measuring the relative algorithmic performances in respect of RM, the ALINEA RM control strategy, which is often hailed as the benchmark RM control strategy (Rezaee *et al* 2013), is also implemented. In ALINEA, the metering rate is adjusted based on the difference in measured and target densities directly downstream of the on-ramp. In the ALINEA control law, a metering rate measured in veh/h is assumed, where $\rho_t$ denotes the target density and $\rho_{ds}(t)$ denotes the measured downstream density during time period $t$. Furthermore, PI-ALINEA is also implemented for additional comparative purposes. The metering rate to be applied following the PI-ALINEA control rule is

$$r(t) = r(t - 1) + K_p[\rho_t - \rho_{ds}(t)] + K_t[\rho_t - \rho_{ds}(t)]$$

Finally the metering rate, taking into account the on-ramp queue consideration limit, may be calculated as

$$r'(t) = \frac{1}{T} [w^* - w(t)] + d(t - 1),$$

where $T$ denotes the length of each control interval $t$, $w(t)$ denotes the measured on-ramp queue length, $w^*$ denotes the maximum allowable on-ramp queue length, and $d(t - 1)$ denotes the on-ramp demand during the previous time period $t - 1$. The final metering rate to be applied is then given by

$$r''(t) = \max[r(t), r'(t)],$$

for both ALINEA and PI-ALINEA. The red phase time to be applied in the microscopic traffic simulation model is then determined from the metering rate as

$$R(t) = \max \left[0, \frac{3600}{r''(t)} - G(t) \right],$$

where $G(t)$ denotes the fixed green phase duration applied at the on-ramp.

We consider five cases in which only RM is applied. In the first case, to serve as a benchmark, no control is applied, while in the second case, RM is enforced according to the modified ALINEA control law in (4). In the third case, the PI-ALINEA control law is adopted, while Q-Learning agents are...
employed to control the RM in the fourth case, and kNN-TD learning agents are employed for this purpose in the fifth case.

The R300 RM agent receives information on the downstream density $\rho_{ds}$ at the section of highway directly downstream of the on-ramp where vehicles joining the highway from the on-ramp enter the highway traffic flow. The upstream density $\rho_{us}$ is measured on the section of highway between the R300 off-ramp at D1 and the R300 on-ramp at O2, while the queue length $w$ is the number of vehicles present in the R300 on-ramp queue.

The downstream density for the Brackenfell Boulevard RM agent is again measured at the section directly downstream of the on-ramp where the traffic flows from the on-ramp and the highway merge. The upstream density is measured on the section of highway between the R300 on-ramp at O2 and the Brackenfell Boulevard on-ramp at O3. Finally, the queue length is again the number of vehicles present in the on-ramp queue.

Similarly, for the Okavango Road RM agent, the downstream density is measured at the section where the on-ramp and highway traffic flows merge, while the upstream density is measured on the section of highway between the Okavango Road off-ramp at D2 and the Okavango Road on-ramp at O4. Finally, as was the case for both the other RM agents, the queue length is the number of vehicles present in the on-ramp queue.

An empirical parameter evaluation was performed to find the best-performing combination of on-ramps at which to employ RM in the case study area, as well as to determine the best-performing target densities for the RM agents at each on-ramp. In the case of ALINEA, PI-ALINEA and Q-Learning, having only one RM agent at the Okavango Road on-ramp, with a target density of 31.2 veh/km, 28.8 veh/km and 31.6 veh/km respectively, yielded the best performance. Furthermore, setting the value of $K_R$ in (4) to 40 yielded the most favourable results for ALINEA. The controller parameters $K_p$ and $K_c$ in the PI-ALINEA implementation were set to 60 and 40 respectively, as these values yielded the best performance. For kNN-TD RM, having an RM agent at both the R300 on-ramp and the Okavango Road on-ramp, with target densities of 28.0 veh/km and 35.5 veh/km respectively, resulted in the best performance. Finally, the maximum allowable queue length was set to 50 vehicles at each of the on-ramps. A more detailed presentation of the parameter evaluations and algorithmic implementations may be found in Schmidt-Dumont (2018).

Summaries of the performances of the resulting algorithmic implementations are provided in Figure 7 and Table 3. The values of the aforementioned performance measures were calculated as the average values recorded after 30 independent simulation runs with varying seeds. For the purpose of comparison, however, the same 30 seed values were employed in each of the cases employing different RM agents.

As may be seen in Table 3, all the RM implementations are able to improve on the no-control case in respect of the TTS.

### Table 3: Algorithmic performance results for RM

<table>
<thead>
<tr>
<th>PMI</th>
<th>No control</th>
<th>ALINEA (%)</th>
<th>PI-ALINEA (%)</th>
<th>Q-Learning (%)</th>
<th>kNN-TD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTS</td>
<td>1 960.01</td>
<td>98.37%</td>
<td>94.45%</td>
<td>93.00%</td>
<td>89.30%</td>
</tr>
<tr>
<td>TTSN1</td>
<td>844.11</td>
<td>106.99%</td>
<td>110.21%</td>
<td>107.12%</td>
<td>72.81%</td>
</tr>
<tr>
<td>TTSR300</td>
<td>992.19</td>
<td>93.19%</td>
<td>81.79%</td>
<td>82.99%</td>
<td>106.44%</td>
</tr>
<tr>
<td>TTB</td>
<td>69.71</td>
<td>96.56%</td>
<td>104.30%</td>
<td>105.52%</td>
<td>87.16%</td>
</tr>
<tr>
<td>TTSO</td>
<td>14.00</td>
<td>228.64%</td>
<td>255.57%</td>
<td>148.57%</td>
<td>125.86%</td>
</tr>
</tbody>
</table>
Interestingly, kNN-TD RM is the only implementation that resulted in a reduction of the TTSN1. This may be explained by the fact that the highway flow is more protected as a result of RM at two of the three on-ramps, while vehicles experience congestion at the R300 on-ramp in all other implementations. As may have been expected, kNN-TD RM resulted in an increase in the TTSR300, as the flow of vehicles from the R300 is metered by RM at the Okavango Road on-ramp, namely VSLR (which again corresponds to VSL_app in Figure 3), is applied. This speed limit is enforced until directly after the Okavango Road off-ramp, which leads to D2. After the off-ramp at the Okavango Road interchange, VSLR (which corresponds to VSL_app in Figure 3) is applied up to the section directly after the Okavango Road on-ramp, at which point the normal speed limit of 120 km/h is restored.

Similarly to the RM implementations, a feedback controller was implemented as a performance benchmark for the RL implementations. The chosen feedback controller is the so-called mainline traffic flow controller (MTFC) by Müller et al. (2015). The control structure of this controller is similar to that of ALINEA, as a VSL metering rate

\[ b(t) = b(t - 1) + K_f [\rho^* - \rho_{ds}(t)] \tag{9} \]

is assumed, where \( K_f \) denotes the controller parameter, and \( \rho^* \) and \( \rho_{ds} \) again denote the target and measured downstream density, respectively. The VSL to be applied at the section VSL_app in Figure 3 is then given by

\[ \text{VSL} = 20 + 100b(t) \tag{10} \]

rounded to the nearest 10 km/h, resulting in speed limits

\[ \text{VSL} \in (20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120). \]

As for the RM implementations, a parameter evaluation was performed for VSLs in order to determine the best-performing combination of VSL agents in the case study area, as well as the best-performing \( K_f \)-value in (9) and \( \delta \)-value in (3) for updating VSL_app, and VSL_O which respectively. This parameter evaluation revealed that the best performance is achieved with an MTFC controller at the Okavango Road interchange with a \( K_f \)-value set to 0.005 and target density set to 37 veh/km. The best performance for Q-Learning was achieved having a VSL agent at the R300 interchange with \( \delta = 10 \), and a VSL agent at the Brackenfell Boulevard interchange. For kNN-TD VSL the best performance is achieved for a VSL agent at the R300 interchange with \( \delta = 20 \) at the R300 interchange, and a VSL agent at the Okavango Road interchange with \( \delta = 10 \). Summaries of the resulting algorithmic performances are provided in Figure 8 and Table 4.

As may be seen in the Table 4, all three VSL implementations were able to achieve reductions in the TTS. In the case of the RL implementations, these reductions are a result of reduced travel times for vehicles travelling along the N1 and entering the N1 from the R300 on-ramp. This may have been expected as these are the vehicles that spend...
the longest time on the N1, where VSLs have the largest effect. The reduction in the TTSR further suggests that effective applications of VSLs may improve the process whereby vehicles from the R300 join the N1 highway. Finally, the reduced variances, as may be seen in Figure 8(b), suggest that there may be successful homogenisation of traffic flow on the N1 due to VSLs. The improvements in respect of the MTFC controller were achieved mainly by those vehicles joining the N1 from the R300, as these vehicles experienced the benefit of VSLs at the Okavango Road interchange, while the effectiveness for vehicles travelling along the N1 only was limited, due to these vehicles still experiencing congestion at the R300 on-ramp merge.

**MARL for RM and VSLs**

As a benchmark for the MARL implementations, the integrated feedback controller of Carlson et al (2014) is implemented. RM occurs according to the PI-ALINEA control law with the addition of a queue limit as in (5), while the MTFC controller of Müller et al (2015) is employed for the control of VSLs. Due to the finding that both PI-ALINEA and MTFC were most effective at the Okavango Road interchange, only one integrated controller is implemented at this interchange.

Due to the fact that the $k$NN-TD learning algorithm achieved the largest reductions in the TTS in both the single agent implementations, only the $k$NN-TD algorithm is implemented in the three MARL approaches. For both the RM and VSL implementations, the best results were achieved by having two RM or VSL $k$NN-TD RL agents in the case study area. The first of these is at the R300 interchange, while the second is at the Okavango Road interchange. As a result, only these two locations are considered for the MARL implementations. The first MARL implementation corresponds to the R300 interchange and consists of the ramp meter placed at the R300 on-ramp, denoted by $O_2$, and the speed limits $VSL_{R}$ and $VSL_{L}$. The target density of the agents in this MARL implementation is set to 28 veh/km, which was determined to be the best-performing target density in the RM parameter evaluation at the R300 on-ramp. $VSL_{R}$ is updated with $\delta = 20$, which was found to yield the best results in the VSL parameter evaluation conducted for VSLs at the R300 interchange.

The second MARL implementation controls the ramp meter placed at the Okavango Road on-ramp, denoted by $O_4$, and the speed limits $VSL_{O}$ and $VSL_{O}$. The target density of the agents in the second MARL implementation is set to 35.5 veh/km, which was determined to be the best-performing target density in the RM parameter evaluation. Finally, $VSL_{O}$ is updated with $\delta = 10$, which was found to yield the best performance in the VSL parameter evaluation. Summaries of the results achieved by the MARL implementations are provided in Figure 9 and Table 5.

As may be seen in Table 5, employing MARL in order to solve the RM and VSL problems simultaneously may lead to further reductions in respect of the

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No control</th>
<th>MTFC</th>
<th>Q-Learning</th>
<th>$k$NN-TD</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTS</td>
<td>1 960.01</td>
<td>95.09%</td>
<td>93.55%</td>
<td>93.92%</td>
</tr>
<tr>
<td>TTSN1</td>
<td>844.11</td>
<td>111.58%</td>
<td>97.04%</td>
<td>103.91%</td>
</tr>
<tr>
<td>TTSR300</td>
<td>992.19</td>
<td>84.62%</td>
<td>93.99%</td>
<td>81.61%</td>
</tr>
<tr>
<td>TTSBB</td>
<td>69.71</td>
<td>96.30%</td>
<td>96.90%</td>
<td>100.63%</td>
</tr>
<tr>
<td>TTSO</td>
<td>14.00</td>
<td>102.50%</td>
<td>101.86%</td>
<td>102.57%</td>
</tr>
</tbody>
</table>

**Table 4 Algorithmic performance results for VSLs**

![Figure 9](https://via.placeholder.com/150)

**Figure 9** Performance measure indicator results for the no-control case (NC), the integrated feedback controller (Feed.), independent MARL (Indep.), hierarchical MARL (Hier.) and maximax MARL (Maxi.)
TTTS, when compared with the single-agent implementations. Furthermore, the combination of RM and VSLs may lead to improved homogenisation of traffic flow, as may be deduced from the smaller variances of the box plots corresponding to the MARL approaches in Figure 9. Finally, the MARL implementations were again able to outperform the integrated feedback controller, providing an illustration of the effectiveness of the MARL approach when compared with the current uncontrolled situation, as well as with the state-of-the-art feedback controller.

CONCLUSIONS

The results obtained from the RM implementations demonstrate that RM may effectively be employed to reduce the total travel time spent by vehicles in the system by up to 10.70% when compared with the no-control case in the context of the case study. Furthermore, the RL approaches to RM outperformed ALINEA and PI-ALINEA, which achieved reductions of only 1.63% and 5.55% respectively. Furthermore, $\text{NN-TD}$ learning was able to find the best trade-off between balancing the length of the on-ramp queue and protecting the highway traffic flow. $\text{NN-TD}$ RM was also the only implementation able to reduce the TTS when an RM agent is present at the R300 on-ramp.

Although they were not quite as effective as RM in reducing the TTS, the VSL implementations resulted in significant reductions in the TTS of 6.45% and 6.08% by Q-Learning and $\text{NN-TD}$ VSL respectively, while the MTFC controller achieved a reduction of 4.91% when compared with the no-control case. One reason for this may be the homogenisation of traffic flow, as the traffic flow becomes more stable at lower speeds, while the results suggest that the process by which vehicles join the N1 from the R300 on-ramp also occurs more smoothly if VSLs are employed in an effective manner.

Finally, employing a MARL approach to solving the RM and VSL problems simultaneously has shown that further reductions in the TTS are possible when these control measures are employed together, as independent MARL, hierarchical MARL and maximax MARL achieved reductions in the TTS of 10.13%, 12.70% and 10.39% respectively over the no-control case. The integrated feedback controller, on the other hand, was only capable of achieving an improvement of 3.36%. Notably, the hierarchical MARL and maximax MARL approaches were able to find the most effective balance between managing the on-ramp queue at the Okavango Road on-ramp and protecting the highway traffic flow, as both of these implementations did not result in statistically significant increases in respect of the TTTSO. Based on these results, we believe there is a strong case to be made in respect of considering the adoption of hierarchical MARL for RM and VSLs on South African highways within urban areas, as significant reductions in the travel times experienced by motorists may be expected.

ACKNOWLEDGEMENTS

The authors would like to thank Mrs Megan Bruwer from the Stellenbosch Smart Mobility Laboratory within the Department of Civil Engineering at Stellenbosch University for her assistance in obtaining the data required for this study.

LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>PMI</th>
<th>Algorithm</th>
<th>No control</th>
<th>Feedback</th>
<th>Independent</th>
<th>Hierarchical</th>
<th>Maximax</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTTS</td>
<td></td>
<td>1960.01</td>
<td>96.64%</td>
<td>89.87%</td>
<td>87.30%</td>
<td>89.61%</td>
</tr>
<tr>
<td>TTSN1</td>
<td></td>
<td>844.11</td>
<td>107.11%</td>
<td>72.83%</td>
<td>73.74%</td>
<td>102.69%</td>
</tr>
<tr>
<td>TTSR300</td>
<td></td>
<td>992.19</td>
<td>87.55%</td>
<td>105.44%</td>
<td>102.36%</td>
<td>81.62%</td>
</tr>
<tr>
<td>TTSBB</td>
<td></td>
<td>69.71</td>
<td>104.09%</td>
<td>84.34%</td>
<td>82.87%</td>
<td>91.45%</td>
</tr>
<tr>
<td>TTSO</td>
<td></td>
<td>14.00</td>
<td>342.14%</td>
<td>289.86%</td>
<td>97.29%</td>
<td>109.71%</td>
</tr>
</tbody>
</table>

REFERENCES


Journal of the South African Institution of Civil Engineering
ISSN 1021-2019
Vol 61 No 3, September 2019, Pages 20–31, Paper 0216

Implementation of a performance-grade bitumen specification in South Africa

S J Bredenhann, P A Myburgh, K J Jenkins, J S O’Connell, G M Rowe, J D’Angelo

South Africa has been experiencing higher traffic volumes and heavier loads over the past several years. This has been accompanied by an increase in premature asphalt failures. Selection of the appropriate asphalt binder is critical for improving performance. Empirical bitumen testing has increasingly failed to relate test results to pavement performance, as the traffic volume and loading have changed. Moreover, empirical tests cannot effectively characterise polymer-modified bitumens that are increasingly being used in South Africa. This changing environment calls for the establishment of specifications based on fundamental engineering properties which relate to actual pavement performance. This paper discusses the fundamental principles of the performance grade (PG) specification being introduced in South Africa. It explains how these fundamental principles create a rational framework for the specification, and present a clear set of compliance criteria to ensure the optimal selection of bituminous binders. The reasons for transitioning to a PG specification are discussed, along with the basis of the specification, rheological concepts, measurements required to characterise bituminous binders, and the simulation of ageing in relation to durability. The framework of the specification, including test procedures, provisional compliance limits and mandatory reporting of test results as an interim measure, are discussed in this paper. Finally, some test results received to date for selected typical South African binders are reported and evaluated. These results indicate that current binders being used in South Africa can easily conform to the proposed PG specification in terms of deformation requirements without any disruption to the processes of the refineries or secondary manufacturers. The fatigue properties of binders are still under evaluation and thus not included in the specification framework. Only information gathered will inform the final decision.

INTRODUCTION

The bituminous product industry endeavours to ensure quality and durability in pavement construction by pursuing mixture tests to characterise bituminous materials in pavements. Tests on bituminous layers (e.g. asphalt) are, however, not feasible for defining the binder requirements. The bituminous binder is a key component in a bituminous layer, and performance-related test methods must be developed to simulate and consider environmental factors to which the binder will be subjected over its lifetime.
Performance-related binder specifications should therefore provide a means to select an appropriate binder in an economic, unbiased and competitive way, and should classify bituminous binders on unbiased grounds, while considering possible changes in the binder rheology experienced during the production and the service life of the asphalt layer (Smith et al 2018). The performance grade specification for South Africa was initiated at the CAPSA Conference in 2004 (Van de Ven et al 2004).

**BACKGROUND**

Recent developments in South Africa, such as the proposed implementation of the SANRAL-sponsored SA Road Design System (SARDS) and the SABITA-sponsored revision of a national asphalt mix design method – introduced to practice during 2016 as SABITA Manual 35/TRH 8 – necessitated the adoption of a specification framework for bituminous binders based on engineering properties to ensure the optimal performance of flexible pavements, especially in the higher traffic categories.

In the USA the implementation of such a specification type in 1995, termed a performance grade (PG) system, was deemed to have yielded notable benefits such as:
- Test conditions suited to specific environmental conditions of climate and traffic
- The introduction of the measurement of the rheological properties of bituminous binder, which gave a better understanding of their behaviour in a range of operating conditions
- The importance of assessing long-term ageing characteristics
- Market shifts to accommodate regional requirements in terms of binder grades
- A rational understanding of when to modify straight-run bitumen.

In developing a PG specification for bituminous binders in South Africa, it is recognised by the Road Pavement Forum (RPF) Task Group on PG specifications, that the performance parameters and compliance limits developed in the USA are based on the requirements of binder performance in asphalt layers, while a significant proportion of binders used in South Africa are applied in chip seals. As it is not feasible, given the market size, to develop distinct specifications for asphalt and seals, the decision was taken to focus on the requirements for asphalt, given the extensive progress in developing such a specification. However, the task group is mindful that, with time, some adjustments or amendments may have to be effected to the proposed specification to cater specifically for the performance requirements of binders used in chip seals. Having said that, it should be noted that there should be compelling reasons for doing so, in order not to render the specification unnecessarily complex, given the size of the bituminous product market in South Africa.

**CURRENT SPECIFICATION**

The current South African national standard specification (SANS 4001-BT1), which can best be described as a nominal grading system, has distinct limitations in meeting the challenges associated with the technological developments mentioned above, such as:
- It merely defines the “consistency” (or viscosity) of the binder over a range of temperatures (from 25°C to 135°C) in terms of surrogate properties such as penetration, softening point and dynamic viscosity.
- Compliance limits refer to neat and short-term aged binder which emulates ageing of the binder, e.g. during mixing and paving of asphalt; there is no assessment of the in-service longer-term ageing characteristics to assess the durability of the binder.
- The specification framework does not cover critical factors related to:
  - Traffic levels and speeds
  - Climate (especially temperature ranges)
  - Critical performance characteristics such as excess viscous flow during periods of elevated temperature or crawling traffic or cracking of the binder due to traffic loading and stresses and strains originating from daily pavement temperature fluctuations, especially as the binder stiffness increases with ageing.

As a direct result of these limitations, the basis of selection of binder grades or types for specific applications is often arbitrary or conventional.

**EVOLUTION OF RHEOLOGICAL TEST METHODS**

The evolution of rheological test methods is summarised in Figure 1. Penetration and softening point tests allow for evaluation of binders in terms of “hardness” or “softness”. The viscosity measurements that followed used kinematic and dynamic methods to evaluate the viscosity of the binder.
The following era brought the advent of advanced dynamic shear rheometry and bending beam rheometry that provided for control over shear strain levels and rates, temperature controls and a variety of loading configurations, as well as the proportions of the elastic and viscous components. Subsequently, testing has expanded into damage beyond the linear-visco-elastic (LVE) range, opening new avenues of deformation and fracture evaluation, primarily at a research level.

DEFINITIONS, CONCEPTS AND TESTING IN PROPOSED SPECIFICATION

Bituminous layers in service are required to resist the following forms of distress:

- Excessive visco-plasticity which may result in permanent deformation and/or bleeding
- Fracture due to the following factors, singly or in conjunction with each other:
  - Fatigue from repeated traffic loading
  - Fatigue induced by daily pavement temperature fluctuations
  - Low temperature cracking
  - Loss of pliability due to ageing.
Asphalt layers should also be durable, i.e. maintain their properties to counter distress over a long period.

As a result, the concept of damage resistance characteristics (DRC), developed in 2010 in the USA, was introduced to provide a specification framework to gauge the binder’s resistance to damage resulting from:

- Permanent deformation (viscous flow) – at elevated temperatures and slow rates of loading
- Cracking – at intermediate temperatures
- Temperature fracture – at low temperatures, that would also address fatigue.

The specification framework should facilitate rational selection of binders based on:

- Traffic volumes and speed
- Climate (maximum and minimum temperatures)
- Binder durability.
Further to the above performance-related properties, additional requirements need to be set in the interests of safety, storage stability and viscosity limits to facilitate handling and application at elevated temperature.

Behaviour of bituminous binders

Bituminous binders display both elastic and viscous behaviour, depending largely on temperature and loading duration or frequency. This visco-elastic character results in a variable response behaviour under varied loading times and temperatures changes.

Elastic behaviour

This behavioural characteristic can be divided into three categories:

- At low temperature and short duration loads bitumen tends to act as an elastic solid, returning to its original position after removal of the load.
- Excessively low temperature in conjunction with rapid loading may cause brittle failure and cracking.
- Low temperature can cause a build-up of internal stress resulting in thermal fracture.

Viscous behaviour

At elevated temperatures and/or low frequency loads associated with slow-moving traffic, bitumen acts as a viscous fluid. It will undergo plastic deformation, i.e. the deformation is not fully recovered. Pavement layers bound with bituminous materials will tend to deform (i.e. rut or flow) under repeated applications of wheel loads, depending on the temperature and rate of loading. It is important to note, though, that this viscous behaviour of the bitumen at high temperatures can be offset by the interlocking action of the aggregate in asphalt mixtures, which serves to resist permanent deformation.

Flow in the binder takes place as adjacent bitumen molecules slide past each other, the resulting friction or resistive force being related to the relative velocity of sliding. The relationship of this resistive force and the relative velocity (of sliding) is termed “viscosity”.

Dynamic viscosity, \( \eta \), is thus a measure of the resistance to flow of a fluid and is expressed as:

\[ \eta = \frac{\text{Resistive force}}{\text{Relative velocity of sliding}} \]

The SI unit of dynamic viscosity is Pascal second (Pa.s)

Visco-elastic behaviour

When a bituminous binder is subjected to constant stress, an instantaneous strain and elastic response to the applied stress takes place, followed by a gradual increase in deformation (or strain) until the load is removed. This ongoing deformation is caused by the viscous behaviour of the material. Upon removal of the stress, the elastic strain is recovered instantaneously and some additional recovery occurs with time – known as delayed elasticity. Ultimately a permanent residual deformation remains, which is non-recoverable and is directly associated with the viscous behaviour of the binder.

Laboratory requirements

In the interests of resource economy, the aim was to limit specialist laboratory equipment to:

- Dynamic Shear Rheometer (DSR)
- Bending Beam Rheometer (BBR)
- Rolling Thin Film Oven (RTFO)
- Pressure Ageing Vessel (PAV).

Dynamic shear rheometer

Measurements using the DSR are the cornerstone of performance grade specifications. It illustrates important components of viscoelastic behaviour and is currently being introduced in South Africa for specification purposes.
Two categories of measurement are afforded by the DSR:
- Fundamental rheological properties
- Creep behaviour of a bituminous binder.

**Fundamental rheological properties**
Both viscous and elastic behaviour is assessed by measuring the complex shearing modulus, \( G^* \) (G-star) and the phase angle, \( \delta \) (delta). \( G^* \) is a measure of the total resistance of a material to deformation when exposed to shear load pulses. The phase angle, \( \delta \), indicates the relative proportions of recoverable and non-recoverable deformation.

When testing within the linear visco-elastic range of the bituminous binder:

\[
G^* = \frac{\text{Maximum applied shear stress}}{\text{Maximum resulting shear strain}}
\]

The phase angle, \( \delta \), represents the time lag between the maximum applied shear stress and the maximum resulting shear strain as seen in Figure 2.

**Creep behaviour**
Creep behaviour of bitumen (in asphalt mixes especially) is very important. If an asphalt layer is loaded there will be an immediate elastic response followed by a delayed elastic / viscous response as long as the load is applied. As the load is removed, part of the elastic deformation will recover immediately, followed by a delayed elastic recovery with a permanent deformation due to the unrecovered deformation resulting from the viscous response under load.

**Definition of minimum, maximum and intermediate temperatures**
Three categories of temperature are adopted to take account of climate and operating conditions in developing a compliance framework. These are:
- Maximum pavement temperature
- Minimum pavement temperature
- Intermediate pavement temperature.

**Maximum pavement temperature \( (T_{\text{max}}) \)**
The maximum pavement temperature is the average annual seven-day maximum at 20 mm depth in the asphalt layer. The maximum pavement temperatures are adjusted to reflect a 98% statistical reliability using climatic data from a minimum of 20 years. The climatic data was converted to pavement temperatures using ThermalPads software developed by the CSIR. The ThermalPads software uses temperature prediction algorithms developed by Viljoen (2001), which are considered more accurate and consistent than the original SUPERPAVE® algorithms (Denneman 2007). The high pavement temperatures were used by the Kriging interpolation method to create a high pavement temperature map for South Africa (O’Connell et al 2015). Figure 3 illustrates the resultant isotherms, with the main isotherm bands being based on the SUPERPAVE® grading classification. Figure 3 shows that South Africa can be divided into two seven-day mean maximum temperature zones (separated by thick red line), the north-western half being 64°C max and the south-eastern half being 58°C max.

**Minimum pavement temperature \( (T_{\text{min}}) \)**
A similar process to that described above was adopted to determine the minimum pavement temperatures in South Africa. Figure 4 indicates that the minimum temperature in South Africa rarely falls below –10°C. However, for grading purposes, considerably lower minimum temperatures were adopted for South Africa, in order to take the following into consideration:

Should actual minimum pavement temperatures be adopted in South Africa, testing associated with the minimum temperatures would not be possible, as sample temperatures would deform at such high minimum temperatures.
It was decided to maintain a temperature difference of 80°C between the maximum and minimum temperature grades in order to have good durability for South African binders.

**Intermediate pavement temperature (T_{it})**
The intermediate pavement temperature is defined as the mid-temperature between the maximum and minimum temperatures plus 4°C, thus:

\[ T_{it} = \frac{T_{\text{max}} + T_{\text{min}}}{2} + 4 \]

**Traffic classification**
Traffic in the PG specification is classified both in terms of volume (or severity) and speed. This is done to take account of the fact that, for a given loading intensity, slow-moving traffic would exert more severe loading conditions. It is proposed that four levels of traffic loading be adopted, in terms of E80s and ruling speed to provide a basis for bitumen selection. Speed categories are chosen to reflect South African conditions and may differ slightly from USA definitions.

Currently it is proposed that the combined effect of traffic loading and speed will be categorised in accordance with Table 1. In this table the following nomenclature is used to describe the severity of the traffic loading:
- **S** refers to Standard conditions
- **H** refers to Heavy conditions
- **V** refers to Very heavy conditions
- **E** refers to Extreme conditions

**Bitumen ageing**
Bitumen in a pavement is subjected to short-term ageing (STA) and long-term ageing (LTA). STA represents the ageing that the binder in an asphalt mix undergoes during the manufacture, storage, transport and placement of the asphalt mix. STA is a high-temperature process which would favour high-energy oxidation mechanisms. LTA represents the ageing that the binder undergoes after placement, and over the lifetime of the pavement. However, for the purposes of defining LTA with regard to correlation with laboratory ageing, and especially with specification purposes in mind, LTA can be defined as the ageing a pavement layer undergoes during a period of five to ten years after placement. LTA is a medium-temperature process ranging between -5°C and 70°C in South Africa. It would favour lower energy oxidation mechanisms, and volatilisation would not be a major factor, unless the bituminous binder has been cut back with solvents or oils (O’Connell & Steyn 2017).

Ageing impacts pavement performance significantly. Ageing of bituminous binders manifests as an increase in bitumen stiffness, thus impacting on the fatigue life of asphalt mixes and surfacing seals. However, ageing can also have a positive effect by increasing the pavement bearing capacity due to the stiffer material. Various ageing methods are available and have been evaluated by researchers (Ekrens et al 2016; Glover et al 2009; Glover et al 2005; Airey 2003).

No ageing test can simulate the actual in-field performance accurately. In the proposed South African specification, the Rolling Thin-Film Oven (RTFO) test was selected to simulate STA and the Pressure Ageing Vessel (PAV) test was selected to simulate LTA for implementation in the new South African PG Binder Specification. These two tests are considered practical to implement. In a recent study Smith et al (2018) concluded on PAV: “Based on the PAV rate at depth, the current R28 protocol of 20 hours seems on track to simulate 5 to 10 years of aging for the binder that was placed at depth.”

**Stiffness and relaxation parameters**
In the decades preceding the SHRP research programme considerable attention was given to the relationship between ductility and pavement performance. An excellent insight into the summary of the pre-SHRP state-of-the art is described in the papers that are part of an ASTM Symposium on the Low-Temperature Properties of Bituminous Materials and Compacted Bituminous Mixtures. It was generally acknowledged that ductility is an empirical property, but it was also associated by early researchers with temperature dependency and shear rate dependency (Vallerga & Halstead 1971). Kandhal (1977) reported on ductility measurements performed at 4°C and 15.6°C and showed a strong relationship between the ductility at 15.6°C (5 cm/min) and progressive pavement deterioration, as reflected in the loss of fines, ravelling and cracking. Poor pavement performance was therefore related to a limiting stiffness (penetration at 25°C) and a limiting ability to relax stresses (ductility at 15.6°C). Glover et al (2005) performed an extensive study on the ageing of asphalt binders directed at “developing an improved method of screening asphalt binders for long-term pavement performance”. Building upon Kandhal’s (Kandhal 1977) success with 15°C ductility, Glover developed a DSR function, \( G'(\eta'G') \) as a surrogate for 15°C ductility, by consideration of a mechanical model that would best describe the ductility test, and then assessing what parameters would best describe the mechanical model with rheological tests. The function, suitable for measurements with the DSR, requires less material than the ductility test and was further verified by performing tests to establish a correlation between the function \( G'(\eta'G') \) measured at 15°C and 0.005 rad/s and ductility at 15°C. Glover et al also suggested that, using time-temperature superposition, the test could be performed at the more user-friendly conditions of 44.7°C and 10 rad/s (Glover et al 2005). The DSR properties were measured on residue from a modified PAV procedure that included thinner films and a longer ageing time. Excellent correlations were found between pavement performance and the DSR function for conventional binders. The use of the viscosity function was complex for modified binders, and no general correlation could be found. Some key aspects noted

### Table 1 Traffic classification

<table>
<thead>
<tr>
<th>Design traffic (million E80)</th>
<th>Traffic speed (km/h)</th>
<th>Asphalt mix design level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt; 20</td>
<td>20 - 80</td>
</tr>
<tr>
<td>&lt; 0.3</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>0.3 – 3</td>
<td>H</td>
<td>S</td>
</tr>
<tr>
<td>&gt; 3 – 10</td>
<td>V</td>
<td>H</td>
</tr>
<tr>
<td>&gt; 10 – 30</td>
<td>E</td>
<td>V</td>
</tr>
<tr>
<td>&gt; 30</td>
<td>E</td>
<td>E</td>
</tr>
</tbody>
</table>
in the Glover report (Glover et al 2005) included:

- Literature reports indicate that the ductility of binders recovered from asphalt pavements correlate with cracking failure. However, ductility measurement is a time- and material-consuming process, and is subject to reproducibility difficulties, as are all failure tests.
- From an elongation model using a Maxwell element, it was seen by Glover et al (2005) that two rheological parameters are suggested to represent the extensional behaviour of asphalt binders in a ductility test – the ratio of the dynamic viscosity to the storage modulus (q'/G') and the value of the storage modulus G'.
- For conventional asphalt the function G'(q'/G') can serve as a surrogate for ductility at 15°C, is easier to measure, and requires less material.

The Superpave PG Grading specification, AASHTO M 320, specifies a minimum stiffness and minimum m-value at the grading temperature. The Strategic Highway Research Program validation report (Leahy et al 1994) contains a plot of the BBR stiffness versus the m-value with four quadrants indicating relative pavement performance. Leahy et al (1994) concluded that “neither parameter is solely responsible for rejection of these binders”, verifying the inclusion of both the stiffness and the m-value in the PG grading specification. Their conclusion is the stiffness and the m-value in the PG binders”, verifying the inclusion of both solely responsible for rejection of these pavement performance. Leahy (1994)

Highway Research Program validation temperature. The Strategic stiffness and minimum m-value at the AASHTO M 320, specifies a minimum

The Superpave PG Grading specification, AASHTO PP78-16 (Design Considerations for PG Graded Asphalt Binders) and AASHTO PP78-16 (Design Considerations When Using Reclaimed Asphalt Shingles (RAS) in Asphalt Mixtures) and is calculated as follows:

\[ T_s = T_1 + \frac{\log (300) - \log (S_1)}{\log (S_1) - \log (S_2)} - 10 \] (1)

\[ T_m = T_1 + \frac{0.300 - m_1}{m_1 - m_2} - 10 \] (2)

\[ \Delta T_m = T_s - T_m \]

Where:

- \( T_s \) = temperature at which the stiffness is 300 MPa
- \( T_m \) = temperature where the m-value is 0.300
- \( T_{1s}, T_{2s} \) = two adjacent specification grading temperatures
- \( S_{1s}, S_{2s} \) = stiffness measurements at \( T_{1s}, T_{2s} \) such that one value passes the specification requirement and one fails the requirement
- \( m_{1s}, m_{2s} \) = m-values at \( T_{1s}, T_{2s} \) such that one value passes the specification requirement and one fails the requirement.

The \( \Delta T_m \) is now being proposed in several specifications due to the ease of its measurement. However, the information collected with this parameter does not show a universal correlation with performance. This may be due in part to it effectively defining the shape of the master curve in a similar manner to the rheological index (R-value). The \( \Delta T_m \) defines the curvature of the \( S(t) \) master curve in the time domain (loading time = 60 seconds in high stiffness region), whereas the R-value is typically used with DSR data to define the curvature of the master curve in the frequency domain in the high stiffness region between 100 kPa to 1 GPa. It can be argued that R defines a temperature dependency aspect, whereas the \( \Delta T_m \) defines a frequency (or time) dependency.

It is intuitively obvious that a shape parameter for a stiffness curve will not be solely adequate for the definition of performance. The reason that the \( \Delta T_m \) shows promise (Reinke 2017), in defining performance, is postulated to be that all the materials judged with this parameter have similar low or intermediate temperature properties. Thus, for an equivalent stiffness material the \( \Delta T_m \) defines the ability to relax stresses which are linked to the formation of durability cracks.

Rowe (2011) provided a discussion on the Anderson et al (2011) paper and proposed that Glover’s parameter be modified to consider only the complex shear modulus (G*) and phase angle (\( \delta \)). The parameter which has been referenced as the Glover-Rowe parameter defines a complex shear modulus versus phase angle relationship within a Black space plot, and the progression of ageing of a binder can be tracked as the stiffness increases. The limits applied in the initial studies have been fixed, using the same data that Glover originally used and are 180 kPa and 600 kPa representing the onset of cracking and significant cracking. These values were developed from the initial studies in Pennsylvania, as reported by Kandhal (1977), and may not be appropriate for all temperature/climatic zones, although the initial studies reported seem to suggest they are generally about the correct level. Regardless, this parameter captures a point on the master curve of stiffness properties and is a hardness value that can assist to anchor the \( \Delta T_m \) values. Consequently, it may be advantageous to use both of these parameters in a specification. Other parameters of interest include the use of the crossover modulus (G_0) and or the crossover temperature at the location where the phase angle is 45°.

Thus, in the analysis of pavements in South Africa, sufficient rheology data is being captured that will allow all of these parameters to be determined and related to field performance, thus enabling specifications to be modified should a more preferable parameter emerge as being significant.

**Permanent deformation parameters**

The Multiple Stress Creep Recovery (MSCR) test is used to assess the creep and recovery potential of the bitumen and is performed in a DSR at the maximum pavement temperature. It is especially significant when considering the permanent deformation (rutting potential) of asphalt layers and the bleeding of spray seals.

The MSCR test uses the well-established creep and recovery test concept to evaluate the binder’s potential for permanent deformation (D’Angelo 2010). Using the DSR, a one-second creep load is applied to the asphalt binder sample. After the one-second load is removed, the sample is allowed to recover for nine seconds. Figure 5 shows typical data for a polymer modified binder under repeated loading. The test is started with the application of a low stress (0.1 kPa) for ten creep/recovery cycles, then the stress is increased to 3.2 kPa and repeated for an additional ten cycles.

From this test the Non-recoverable Creep Compliance (\( J_{NR} \) in kPa\(^{-1}\)) for a bituminous
binder is determined as follows, with reference to Figure 6:

\[
J_{NR} = \frac{\text{Average non-recoverable shear strain}}{\text{Applied stress (kPa)}}
\]

The material response in the MSCR test is significantly different to the response in the existing USA SUPERPAVE® system tests. In the SUPERPAVE® system, the high temperature parameter, \(G'/\sin\delta\), is measured by applying an oscillating load to the binder at very low strain. This is one of the reasons why the existing SUPERPAVE® high temperature parameter does not accurately represent the ability of polymer-modified binders to resist rutting. Under the very low levels of stress and strain present in dynamic modulus testing, the polymer network is never really activated. In the existing Superpave specification the polymer is really only measured as a filler that stiffens the asphalt. In the MSCR test, higher levels of stress and strain are applied to the binder, better representing what occurs in an actual pavement. By using the higher levels of stress and strain in the MSCR test, the response of the asphalt binder captures not only the stiffening effects of the polymer, but also the delayed elastic effects, that rubber-band type of effect (D’Angelo 2010).

The MSCR test also gives valuable insight into the recovery behaviour of the material. The compliance value \(J_{NR}\) from the MSCR test provides the rut resistance and the amount of recovered strain (Percent Recovery) from the test identifies the presence of polymer and also the quality of the blending of the polymer in the binder. The determination of Percent Recovery (%R) is illustrated in Figure 7, whereby:

\[
%R = \frac{\varepsilon_0 - \varepsilon_f}{\varepsilon_0} \times 100
\]

Where:
- \(\varepsilon_0\) = initial strain in the creep stage
- \(\varepsilon_f\) = final strain after recovery.
The proposed specification is presented in Table 2. Two main bitumen grades are envisaged for South Africa, PG58-22 and PG64-16, with the minimum temperatures maintaining the 80°C difference with the maximum grade temperature. The high pavement temperature map was created using data points which were relatively far spaced in places. In the event of higher pavement temperatures prevailing as a result of microclimates not addressed by the isotherm map in Figure 3, provision is made for a PG70-10 bitumen grade, again maintaining the 80°C temperature difference.

Requirements are specified for un-aged, RTFO-aged (short-term ageing) and PAV-aged (long-term ageing) bituminous binders.

**Requirements based on an un-aged bitumen**

Requirements specified, based on an un-aged binder, are related to a quick quality control measure ($G^*\sin \delta$) that relates to

<table>
<thead>
<tr>
<th>Test Property</th>
<th>South African Performance Grades</th>
<th>Test Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max pavement design temperature (°C) ($T_{\text{max}}$)</td>
<td>58S-22</td>
<td>58H-22</td>
</tr>
<tr>
<td>$G^*$ and $\delta$ at $([T_{\text{max}} + T_{\text{min}}]/2 + 4)°C$</td>
<td>Compulsory report only – see detail description of report only item</td>
<td>Report only</td>
</tr>
<tr>
<td>$G^*\sin \delta$ @10rad/s (kPa) @ $T = T_{\text{max}}$</td>
<td>≤ 0.9</td>
<td>ASTM D4402</td>
</tr>
<tr>
<td>Viscosity at 165°C (Pa.s)</td>
<td>≥ 30 sec$^{-1}$</td>
<td>≤ 0.9</td>
</tr>
<tr>
<td>Storage stability at 180°C (% diff in $G^*$ at $T_{\text{max}}$)</td>
<td>≤ 15</td>
<td>ASTM D7175</td>
</tr>
<tr>
<td>Flash Point (°C)</td>
<td>≥ 230</td>
<td>ASTM D92b</td>
</tr>
</tbody>
</table>

**After RTFO Ageing**

<table>
<thead>
<tr>
<th>Test Property</th>
<th>South African Performance Grades</th>
<th>Test Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^*$ and $\delta$ at $([T_{\text{max}} + T_{\text{min}}]/2 + 4)°C$</td>
<td>Compulsory report only – see detail description of report only item</td>
<td>ASTM D7175</td>
</tr>
<tr>
<td>Mass change (% m/m)</td>
<td>≤ 0.3</td>
<td>≤ 1.0</td>
</tr>
<tr>
<td>$J_{\text{RTFO}}$ at $T_{\text{max}}$ (kPa)</td>
<td>≤ 4.5</td>
<td>≤ 2.0</td>
</tr>
<tr>
<td>Ageing ratio [$G^<em>_{\text{RTFO}}/G^</em>_{\text{Original}}$]</td>
<td>≤ 3.0</td>
<td>ASTM D7175</td>
</tr>
</tbody>
</table>

**After RTFO plus PAV Ageing**

<table>
<thead>
<tr>
<th>Test Property</th>
<th>South African Performance Grades</th>
<th>Test Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^*$ and $\delta$ at $([T_{\text{max}} + T_{\text{min}}]/2 + 4)°C$</td>
<td>Compulsory report only – see detail description of report only item</td>
<td>ASTM D7175</td>
</tr>
<tr>
<td>Maximum creep stiffness tested at temperature ($T_{\text{min}}$ + 10°C), MPa</td>
<td>≤ 12°C</td>
<td>≤ 0°C</td>
</tr>
<tr>
<td>Minimum m-value tested at temperature ($T_{\text{min}}$ + 10°C), [m (60s) ≥ 0.300]</td>
<td>≤ 12°C</td>
<td>≤ 0°C</td>
</tr>
<tr>
<td>Δ$T_c$ (°C) = $T_{c,S} - T_{c,m}$</td>
<td>≥ –5</td>
<td>ASTM D7643</td>
</tr>
<tr>
<td>Ageing ratio [$G^<em>_{\text{PAV}}/G^</em>_{\text{Original}}$]</td>
<td>≤ 6.0</td>
<td>ASTM D7175</td>
</tr>
</tbody>
</table>

**Figure 7** Plot showing the determination of the % recovery in the MSCR test
the old penetration parameter, a safety measure (flash point) and production measures in viscosity and storage stability.

\[ G^\ast \sin \delta \]

\[ G^\ast \sin \delta \] at 10 rad/sec and \( T_{\text{max}} \) have been retained in the specification only as a quick indication of binder stiffness for quality control purposes. Whilst the \( G^\ast \sin \delta \) requirement is used as the main parameter of rutting potential in the USA Superpave specification, it is not used for this purpose in the South African specification at all. The rutting requirement in the South African specification is the \( J_{\text{NR}} \) parameter. \( G^\ast \) and \( \delta \) must be reported separately.

**Viscosity**

Viscosity is the material characteristic used to describe the resistance of fluids to flow. A specification requirement is needed to make sure that a bitumen can be pumped at a certain (high) temperature. Viscosity at 163°C at a shear rate greater than 30 Hz is the specification requirement included as an indication of pumpability. Unmodified bitumens normally show Newtonian behaviour, i.e. the viscosity is constant regardless of the shear rate; however, modified bitumens show shear thinning behaviour at an intermediate shear rate range. Figure 8 shows a schematic of the typical viscosity variation over a range of shear rates at which viscosity is measured. At very low shear rates in Figure 8, the viscosity is constant and is defined as the Zero Shear Viscosity (ZSV) or \( \eta_0 \). It is impracticable to test at such low shear rates. In the intermediate shear rate range, the binder displays shear-thinning behaviour and the viscosity reduces with an increase in shear rate until at a certain shear rate the Newtonian viscosity, \( \eta_\infty \), is reached. Research by the CSIR (Mturi et al 2013) has shown that \( \eta_\infty \) is usually reached at shear rates larger than 30 Hz. Thus a test temperature of 163°C was selected to ensure that the viscosity is measured at high shear rate where the viscosity becomes constant and independent of shear rate, as is shown in Figure 8.

**Storage stability at 180°C**

The storage stability test is used to determine the degree of separation for a modified bitumen during storage. This test uses 30 cm high test tubes that allow separation at 180°C for a fixed period and then temperature reduction to 0°C to enable fracture of the test tube to allow separate binder specimens to be retrieved from the top and bottom of the tube, for DSR test evaluation. It is not always guaranteed that \( G^\ast \) from either will be the highest, therefore the storage stability is calculated as \( \frac{G^\ast_{\text{HIGH}}} {G^\ast_{\text{LOW}}} \) (from top/bottom sample). The \( G^\ast \) is measured at \( T_{\text{max}} \).

**Flash point (°C)**

The flash point requirement is retained in the new specification for safety purposes.

**Requirements based on a RTFO-aged bitumen**

Three requirements are specified for the RTFO-aged bitumen – the mass change, non-recoverable compliance \( (J_{\text{NR}}) \) and an ageing ratio.

**Mass change**

The mass change specification from the current SANS BT1 is retained in the new specification with requirements of not exceeding 1.0 % m/m for all other traffic classes.

**Non-recoverable compliance (JNR) from MSCR test**

The non-recoverable compliance \( J_{\text{NR}} \) is done at \( T_{\text{max}} \) with the following requirements:

- S-Class, \( J_{\text{NR}} \leq 4.5 \text{kPa}^{-1} \)
- H-Class, \( J_{\text{NR}} \leq 2.0 \text{kPa}^{-1} \)
- V-Class, \( J_{\text{NR}} \leq 1.0 \text{kPa}^{-1} \)
- E-Class, \( J_{\text{NR}} \leq 0.5 \text{kPa}^{-1} \)

**Ageing ratio**

An ageing ratio requirement is specified for RTFO-aged and PAV-aged bitumens that will give an indication of the bitumen's resistance to ageing. \( G^\ast \) and \( \delta \) are measured at 10 rad/s at the intermediate temperature; an 8 mm parallel plate will be used unless the stiffness \( G^\ast < 100 \text{kPa} \), in which case a 25 mm parallel plate will be used. The ageing ratio is calculated as the ratio between the aged bitumen and the un-aged bitumen, i.e. \( G^\ast_{\text{RTFO}} / G^\ast_{\text{Original}} \) for the RTFO-aged bitumen.

**Requirements based on a PAV-aged bitumen**

Three requirements are specified based on the PAV-aged BBR test – the maximum creep stiffness at 60 seconds loading time, \( S(60) \), the minimum m-value at 60 seconds loading time, \( m(60) \) and the \( \Delta T_c \) parameter. The values are determined for \( T_{\text{min}} \) at two hours loading time. Because two hours were considered to be too long for testing, the principles of time-temperature superposition were used to obtain the same values at a temperature of \( T_{\text{min}} +10^\circ \text{C} \) and 60s loading time.

**Maximum creep stiffness**

A maximum requirement for creep stiffness, \( S(60) \), of 300 MPa at \( T_{\text{int}} \) after 60 seconds of loading time is specified to limit the bitumen stiffness with ageing. It is known that stiff binders, caused by either ageing or cold temperatures, will crack. Limiting the bitumen's creep stiffness will limit cracking to cold temperatures. It is also known that fatigue is highly influenced by an increase in stiffness with ageing. The equivalency between cold temperature stiffness and ageing stiffness is therefore currently seen to be an indicator of binder relaxation properties, and hence also an indicator of fatigue resistance.
An ageing ratio of 0.22, as discussed in detail above.

The specified BBR isotherms for the un-aged, RTFO- and PAV-aged bitumen (LTA) are set out in Table 4, and at the frequency ranges in master curve development: Minimum temperature ranges to be tested.

<table>
<thead>
<tr>
<th>PG Grading</th>
<th>DSR @ IT (°C) Unaged-, RTFO- and PAV-aged binder</th>
<th>BBR @ LT (°C) PAV-aged binder only</th>
</tr>
</thead>
<tbody>
<tr>
<td>58–22</td>
<td>22</td>
<td>−12, −18, −24, −30</td>
</tr>
<tr>
<td>64–16</td>
<td>28</td>
<td>−6, −12, −18, −24</td>
</tr>
<tr>
<td>70–10</td>
<td>34</td>
<td>0, −6, −12, −18</td>
</tr>
</tbody>
</table>

Minimum m-value
A minimum requirement for the m-value, m(60), is specified. The m(60) parameter is an indication of the bitumen’s ability to relax. Increasing the relaxation parameter will ensure that the bitumen will relax after a stress application.

ΔT_c
A requirement for ΔT_c of greater than −5°C is specified. The ΔT_c parameter was discussed in detail above.

Ageing ratio
An ageing ratio of G^PAV / G^Original is also specified for the PAV-aged bitumen (LTA). Having ageing ratios for STA and LTA, bitumens can be assessed in terms of the bitumen’s resistance to ageing.

 DETAILS ON REPORT ONLY ITEM
A compulsory report requirement is included in the specification. This includes:
- A single DSR frequency sweep isotherm for un-aged, RTFO- and PAV-aged bituminous binders at [(T_{max} + T_{min})/2 + 4]°C (see Table 3) at an applicable strain level, as set out in Table 4, and at the frequency values set out in Table 5.
- The specified BBR isotherms for the PAV-aged binder; the temperatures are set out in Table 3.
- With the BBR isotherm of the PAV-aged bitumen at the low temperatures and the respective isotherms of the aged binders, full master curves can be developed. The procedure to follow to produce the master curves is the subject of a separate paper.
- The master curves are used to calculate additional rheological parameters and build a database to investigate additional parameters for inclusion in future versions of the specification.

IMPLEMENTATION PLAN
At the Roads Pavement Forum (RPF) held in May 2016 (RPF 2016) it was decided, for a period of two years starting in July 2016, to implement the proposed Specification Framework for performance-graded bituminous binders in parallel with current binder specifications. Implementation is voluntary and will be done on selected SANRAL roads; other provincial and local government road authorities will also join forces. This implementation period is required to gather data on bitumen in terms of the new specification and build a database of typical bitumen requirements with an eventual aim of verifying the limits of all specification requirements. A draft performance-grade bitumen specification was submitted to the South African Bureau of Standards (SABS) that will be published as a SANS Technical Note, SATS 3208. A full SANS Standard Specification will follow at a later stage.

In the meantime, a few seminars to introduce the specification were held across the country, and two training courses were held – in June 2016 at SANRAL Head Office in Pretoria, followed up in April 2017 in Cape Town at the Stellenbosch University Business School, both jointly sponsored by SANRAL and SABITA.

A proficiency testing programme between the different laboratories was also launched where test results are compared on a regular basis to ensure consistency in testing and quality of results. The committee responsible for the proficiency testing programme will also develop test protocols in addition to the specifications to ensure that identical procedures are used in all participating laboratories.

Report back on progress with the implementation of the performance-grade specification is done every six months at the RPF.

RESULTS
Non-recoverable compliance results (I_{NR}) are reported in Table 6 for a number of typical South African binders. Softening points and penetration are added for reference and to compare to the current South African penetration specification. Reference to “1”, “2” and “3” indicates different suppliers for similar products. The I_{NR} results shown were determined at 64°C. Results for 58°C are approximately half of the values obtained at 64°C. These results are not included in the paper for the sake of brevity.

The range of the results for I_{NR} indicate that the binders currently produced in South Africa would satisfy all the requirements of the proposed PG specification classes with regard to I_{NR}.

Table 3 Temperature ranges in master curve development: Minimum temperature ranges to be tested

<table>
<thead>
<tr>
<th>Plate size (mm)</th>
<th>Strain level</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1%</td>
</tr>
<tr>
<td>25</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 4 Strain levels for master curve development

<table>
<thead>
<tr>
<th>Log basis</th>
<th>Linear basis (rad/sec)</th>
<th>Linear basis (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.6</td>
<td>0.251</td>
<td>0.0400</td>
</tr>
<tr>
<td>−0.4</td>
<td>0.398</td>
<td>0.0634</td>
</tr>
<tr>
<td>−0.2</td>
<td>0.631</td>
<td>0.100</td>
</tr>
<tr>
<td>−0.0</td>
<td>1.00</td>
<td>0.159</td>
</tr>
<tr>
<td>+0.2</td>
<td>1.58</td>
<td>0.252</td>
</tr>
<tr>
<td>+0.4</td>
<td>2.51</td>
<td>0.400</td>
</tr>
<tr>
<td>+0.6</td>
<td>3.98</td>
<td>0.634</td>
</tr>
<tr>
<td>+0.8</td>
<td>6.31</td>
<td>1.00</td>
</tr>
<tr>
<td>+1.0</td>
<td>10.0</td>
<td>1.59</td>
</tr>
<tr>
<td>+1.2</td>
<td>15.8</td>
<td>2.52</td>
</tr>
<tr>
<td>+1.4</td>
<td>25.1</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 5 Frequency range for each temperature in master curve development
**RESEARCH**

In a further effort to verify the limits of all specification requirements in the performance-grade specification, a research project has been launched at the Stellenbosch University (SU) and the Council for Scientific and Industrial Research (CSIR), sponsored by SANRAL and Much Asphalt (Pty) Ltd. Use is also made of the laboratories of Colas and Tosas in the research. This research follows on initial research at the CSIR sponsored by University (SU) and the Council for Scientific and Industrial Research (CSIR), sponsored by USA for development and inclusion in future specifications. The data collection scheme being implemented allows a detailed evaluation of these parameters for the South African climate zones and traffic levels. Results to date indicate that current binders being used in South Africa can readily conform to the proposed PG specification in terms of deformation requirements without any disruption to the refineries or secondary manufacturers.

Should more appropriate parameters emerge for controlling cracking behaviour than those currently being specified, some minor modifications can easily be made within the existing specification framework. It should be noted that the climatic zones for South Africa are not vastly different from some other parts of the world where extensive data collection is taking place and we can build upon the experience being captured in these climatic zones.

**SUMMARY AND CONCLUSIONS**

The current proposed format of the PG-grade specification being implemented in South Africa builds upon those implemented in the USA and the parameters considered to be of key interest in the USA for development and inclusion in future specifications. The data collection scheme being implemented allows a detailed evaluation of these parameters for the South African climate zones and traffic levels. Results to date indicate that current binders being used in South Africa can readily conform to the proposed PG specification in terms of deformation requirements without any disruption to the refineries or secondary manufacturers.

**ACKNOWLEDGEMENTS**

This paper is presented with the approval of the South African National Roads Agency Limited SOC. The contents of the paper reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official view or policies of the South African National Roads Agency Limited SOC. The authors also wish to express their thanks to SABITA who initiated the research that eventually culminated in the specifications discussed in this document.

**REFERENCES**


---

**Table 6 JNR results for typical South African binders**

<table>
<thead>
<tr>
<th>Penetration (10⁻² mm)</th>
<th>70/100 Binder 1</th>
<th>70/100 Binder 2</th>
<th>70/100 Binder 3</th>
<th>70/100 Binder 4</th>
<th>50/70 Binder 1</th>
<th>50/70 Binder 2</th>
<th>50/70 Binder 3</th>
<th>35/50 Binder 1</th>
<th>35/50 Binder 2</th>
<th>35/50 Binder 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>85</td>
<td>81</td>
<td>80</td>
<td>65</td>
<td>62</td>
<td>62</td>
<td>48</td>
<td>46</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>45.4</td>
<td>47.8</td>
<td>45.8</td>
<td>48.0</td>
<td>50.0</td>
<td>48.4</td>
<td>49.0</td>
<td>52.0</td>
<td>51.2</td>
<td>52.6</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>1.8</td>
<td>2.0</td>
<td>1.9</td>
<td>3.0</td>
<td>2.8</td>
<td>2.7</td>
<td>5.2</td>
<td>4.5</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>1.3</td>
<td>1.3</td>
<td>1.2</td>
<td>2.3</td>
<td>2.0</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.5</td>
<td>1.0</td>
<td>0.9</td>
<td>1.9</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Summary**

The current proposed format of the PG-grade specification being implemented...


Reinke, G 2017. The relationship of binder delta %Tc (Δ%Tc) to mixture fatigue. Presentation to the SEAUPG Meeting, Florida, USA.


Monitoring permeability potential of hot mix asphalt via binary aggregate packing principles correlated with Bailey ratios and porosity principles

E Horak, J Maina, P Myburgh, H Sebaaly

Asphalt mix designs tend to optimise the load transfer via aggregate skeletons as main mechanism to provide rut resistance, often to the detriment of durability. Permeability, as a significant durability indicator, is more difficult to measure in the field than in the laboratory. Voids in the asphalt mix have a critical zone where an increase in voids is exponentially linked to permeability. This zone is where voids start to become increasingly interconnected. The aggregate grading envelope characteristics can provide an indication of the interconnectedness of the voids to enhance quality control. New rational Bailey Method Ratios (BMRs) were defined with contiguous aggregate fractions in the numerator and denominator. This allows also for porosity calculation using the Dominant Aggregate Size Range (DASR) method. The Binary Aggregate Packing (BAP) triangle porosity diagrams provide insight into the link between porosity and interconnected voids. The wall and the loosening effects create additional porosity (voids) with increased probability of interconnectedness. Clear threshold zones of interconnected voids can be determined with BAP coarse/ fine mass ratios. The latter is the inverse of the rational BMRs. It allows for simple spreadsheet calculations of porosity and coarse/fine mass ratio as a screening tool for probable permeability via benchmark analysis. Reworked data sets demonstrated how the inverse of BMRs could show potential for interconnectedness of voids and, therefore, permeability propensity.

INTRODUCTION

The aggregate stone skeleton is known to form the core of the flexural and compressive strength resistance to permanent deformation and fatigue cracking of Hot Asphalt Mix (HMA) in asphalt pavement layers. High-level theoretical work on tessellation and discrete element modelling are at the cutting edge of research (Alishibi & El-Saidany 2001; Burstea & Werner 2015). However, more practical approaches are available with the correct fundamental characteristics describing aggregate volumetric packing principles. The Bailey method (Vavrik et al 2001; TRB 2002) is well established in the USA as a logical approach to enhance aggregate packing analysis in asphalt mix design. It is based on the fundamental concept that fine aggregates equal or less than 0.22 of a specific larger aggregate size will be able to fit in the voids in-between these large aggregates. This ensures aggregate on aggregate contact, and therefore structural strength of the aggregate skeleton. The Bailey method is also included in the South African Guideline 35 (SABITA 2016) as a valuable method to assist in designing asphalt mixes, which rely on the aggregate skeleton as main load-bearing mechanism in a paved asphalt layer to provide resistance to rutting. The Bailey method is, however, not well articulated with permeability of the asphalt mix.

Permeability, as a durability indicator, is normally measured under controlled environment in the laboratory during the design phase. However, permeability is more difficult to measure in the field. Indirect measures of permeability are often used, like voids in the mix. Research by Cooley et al (2002) has shown that the effect of nominal aggregate or particle size (NMAS or NMPS), lift thickness, density, etc, all influence voids in the mix, and therefore permeability. The relationship between voids in the mix and permeability shows a range of voids in the mix where...
permeability actually starts and thereafter increases exponentially. This start of permeability is where the voids become interconnected and permeability increases with increase in interconnectedness of the increasing voids.

This field of the uncertainty regarding the interconnectedness of voids is a complex field that deals with well-known concepts like absorption, water saturation, capillary hysteresis, capillary pressure, moisture dispersion, tortuous flow in porous media, phase permeability and apparent permeability (Koponen et al 1996; Cherry & Freeze 1997; Cedergren 1997). Complex modelling of hydraulic conductivity has been used with great effect to monitor phases of permeability in isotropy and anisotropy (Kutay et al 2007). However, the approach here is to limit the discussion to normal (practical) road and pavement engineering aspects that have practical application in quality control / quality assurance (QC/QA) systems.

The tendency to design and specify for a minimum density (e.g. 97% minus voids and a minimum 92% Rice density) and a range of air voids (3% to 6%) is clearly intended to address rut resistance in the first place, but the upper limit of the derived voids in the mix is already close to the known precipice where permeability may increase exponentially. The situation in the field is often that marginal densities are accepted based on statistical methods used for density control, but the actual voids in the mix may have shifted over the proverbial precipice (e.g. 7% or more). Due to uncertainty with regard to the interconnectedness of the voids, permeability may often start within the specified voids range, and the voids will have a higher probability to be fully interconnected at higher voids percentages. A better indicator of interconnectedness of voids via aggregate grading characteristics is needed to monitor permeability potential in asphalt mixes as an addition to normal QC/QA measures.

**BAILEY METHOD AS BASIS OF CALCULATIONS**

**Bailey method definitions and ratios**

The fundamental correctness of the volumetric ratio of the Bailey method and the acceptance world-wide as a guide to enhance good asphalt mix designs (Vavrik et al 2001; TRB 2002) make the Bailey method an ideal reference system in the analysis of grading envelopes. In such analyses, the bituminous binder is seen as the glue that keeps the skeleton matrix together via adhesion and cohesion characteristics like the filler binder ratio or mastic or effective film thickness concepts (Butcher & van Loon 2013), but is not considered in this analysis and discussion regarding the link with permeability. Admittedly, this creates a ‘greyness’ in the data, which implies final analysis for monitoring permeability potential may best be served with a benchmark approach. The Bailey method terminology is used throughout for description of specific aspects of the grading envelope. The definitions of various Bailey control sieves and associated ratios are described as follows (Vavrik et al 2001):

- **Maximum Nominal Particles Size (NMPS)** as per the Superpave definition: “One size larger than the first sieve that retains more than 10% aggregate.” This was changed lately to 15% (SABITA 2016).
- **Half Size (HS)** is defined as the sieve size closest to or equal to half the NMPS.
- **Primary Control Sieve (PCS)** is the sieve size closest to 0.22 × NMPS.
- **Secondary Control Sieve (SCS)** is the sieve size closest to 0.22 × PCS.
- **Tertiary Control Sieve (TCS)** is the sieve size closest to 0.22 × SCS.

Additional aggregate size range descriptors and terminology often used in literature are:

- **Interceptors:** The size range between HS and PCS
- **Pluggers:** In the original Bailey method description (Vavrik 2001) they are defined as all aggregate fractions larger than half size. However, “Pluggers” can be divided into:
  - **Pluggers up to NMPS (PN).** The size ranges between HS and NMPS. Horak et al (2017a&b) referred to this as Pluggers.
  - **Oversize pluggers larger than NMPS (PO).** The oversized aggregate is larger than NMPS. Horak et al (2017a&b) referred to this as Oversize.

Three original Bailey Method Ratios (BMRs) were defined as the ratio between various aggregate fractions. These BMRs are used for gauging strength contribution in the aggregate skeleton. They were originally defined as:

\[
\text{Fine aggregate coarse ratio:} \quad FA_c = \frac{\%TCS}{\%PCS} \\
\text{Fine aggregate fine ratio:} \quad FA_f = \frac{\%TCS}{\%SCS}
\]

**Note:** CA ratio consists of contiguous (following each other) aggregate fractions, while both \( FA_c \) and \( FA_f \) have aggregate fractions that overlap. This distinction will become clearer in the next sections.

**Linking Bailey Method Ratios (BMRs) directly with permeability**

Horak et al (2017a&b) investigated various research papers dealing with the link between aggregate grading and structural strength and permeability. Typically, investigations by Al Mosawe et al (2015) confirmed the structural strength of the asphalt mix often tends not to be dependent on the packing efficiency at the macro level of the aggregate grading (typically the Bailey coarse fraction, e.g. the CA ratio) alone, but surprisingly more so at the middle level of the aggregate grading (midi or rather meso level). The original BMRs tend to lack distinction of various fractions in the midi or rather meso level. Al Mosawe et al (2015) confirmed the potential for an interference/disrupter effect caused by the intermediate aggregate sizes (interceptors). In continuously graded mixes (stock trade of South African HMA), the grading adheres to a Talbot-type grading curve, and therefore fine/coarse aggregate size ratios do not always allow for just filling of the voids between the larger pluggers aggregates, but can in effect also be pushing the larger aggregates apart. Therefore, the coarse portion of the fine aggregates range with the interceptors may often take over the dominant structural contribution, as is known to happen with sand skeleton continuously graded HMA (SABITA 2016).

Al Mosawe et al (2015) therefore developed a more descriptive BMR for these mid or rather meso level aggregate fractions covering fine of the coarse to the coarse of fine aggregate range. They are described in Equations 4 and 5. What is significant is that they have contiguous (successive) aggregate fractions in the numerator and denominator. Komba et al (2019) have subsequently also confirmed the value of such contiguous meso level ratios in compaction and strength gain studies. As noted before, the original fine BMRs defined in Equations 2 (\( FA_c \)) and 3 (\( FA_f \)) do not have such contiguous
aggregate fractions and end up as poor indicators of strength or permeability.

$$\frac{C_f}{F_c} = \frac{(%PCS \ - \ %SCS)}{(%HS \ - \ %PCS)}$$  \hspace{1cm} (4)

$$\frac{F}{C} = \frac{(%PCS)}{(%NMPS \ - \ %PCS)}$$  \hspace{1cm} (5)

Sadasivam and Khosa (2006) determined that the permeability of a structurally competent asphalt mix is influenced by the packing arrangement (retained fraction alone and fraction ratios) at the rather meso level (typical interceptors) and their void infill at the micro level by the fines. This was also emulated and confirmed by Dennenman et al (2007) with typical South African-designed continuously graded mixes. Horak et al (2017a&b) summarised such recent research and analysis of the grading characteristics, as well as BMRs and their correlation with strength and permeability. Horak et al (2017a) confirmed it is largely the fines at the micro level that correlates best with permeability by reworking the data sets of a number of such published permeability studies. Of the original BMRs, it is only the CA ratio which showed some correlation with permeability, whereas FA_c and FA_T were insignificant. The results show an improved understanding of the importance of certain meso-range to micro-range aggregate sizes, but still do not give adequate or reliable indication of the permeability potential. This helped to indicate that there was obvious room for improvement in the aggregate grading envelope analysis and the determination of porosity. Porosity as a volumetric indicator seemed to be the best way to get to an improved understanding of how it relates to permeability.

### POROSITY DETERMINATION OF AGGREGATE RATIOS

#### Dominant Aggregate Size Range (DASR)

The limitations of using the BMRs for permeability monitoring (Horak et al 2017b) created the need to investigate the grading envelope in more detail and establish a closer link with permeability or ‘a handle’ on the interconnected voids. Olard (2015) makes use of the nesting principle to ‘deconstruct’ the grading envelope as three subsets of infill aggregate skeletons in analysis and even asphalt mix designs. In this way macro, midi or rather meso and micro level subsets can be defined (Horak et al 2017a), which certainly helps to understand the contribution of various aggregate fractions at various positions on the aggregate grading envelope. It allows for the visualisation of how the increasingly smaller voids of each skeleton subset is filled in with smaller or finer aggregate combinations in specific or preset volumetric proportions. The mantra of the original Bailey method that the ‘voids are in the fines’ (Vavrik et al 2001; TRB 2002) is confirmed via this three-level aggregate skeleton analysis. This implies finer aggregate filling the coarse (macro skeleton), and then even finer aggregate fractions filling the voids left in the midi or rather meso level, and finally micro level fines filling the resultant voids at the micro level. This nesting concept highlights the concept of contiguous aggregate fractions as the basis for determining porosity of sections of an aggregate grading. The need for ratios to adhere to contiguous aggregate ranges indirectly led Al Mosawe et al (2015) to derive the midi or rather meso level BMRs described by Equations 4 and 5.

Porosity has a more direct relationship with permeability in that permeability is facilitated by the interconnectedness of voids in the asphalt mix. The Dominant Aggregate Size Range (DASR) concept has the same basis as the Bailey method in the size ratios (less than 0.22) that determine the densest possible aggregate packing. However, the DASR method relies on porosity determination (Horak et al 2017a&b) of a range of contiguous aggregate fractions that provide the lowest porosity. Low porosity, below 0.5 to 0.4, implies density of aggregate packing and increased aggregate skeleton strength due to interlock.

Dennenman et al (2007) showed how it is possible to simplify the DASR porosity calculation when limited to a single or combination of two contiguous aggregate fractions. In Table 1 the DASR porosity Equation 6 is defined for a contiguous range (often more than two) and the simplified calculation to determine a single fraction, or contiguous range of aggregate fractions, is defined as Equation 7.

The Dennenman et al (2007) analysis of permeability propensity of various South African continuously graded mixes used only the original BMRs. FA_c and FA_T are insignificant to permeability sensitivity for reasons of lack of contiguousness. Nevertheless, the work by Al Mosawe et al (2015) enabled the understanding that contiguous aggregate fractions can in effect be determined for other aggregate fractions along the whole grading envelope and be able to determine porosity for each such contiguous aggregate combinations. This allowed investigation of the application of BMRs and calculated porosity of single aggregate fractions and DASR contiguous aggregate fractions, and

<table>
<thead>
<tr>
<th>Table 1 DASR and contiguous aggregate fraction porosity equations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equation 6 (Kim et al 2006)</strong></td>
</tr>
<tr>
<td>DASR = ( \frac{V_{ICAGG} + VMA}{V_{TM} - V_{AGG &gt; DASR}} )</td>
</tr>
<tr>
<td>Where:</td>
</tr>
<tr>
<td>DASR</td>
</tr>
<tr>
<td>( V_{ICAGG} )</td>
</tr>
<tr>
<td>( V_{AGG &gt; DASR} )</td>
</tr>
<tr>
<td>( V_{TM} )</td>
</tr>
<tr>
<td><strong>Equation 7 (Dennenman et al 2007)</strong></td>
</tr>
<tr>
<td>( \eta(4.75 – 2.36) = \left[ \frac{PP_{2.36}}{100} \right] \left( \frac{V_{TM} - VMA + VMA}{V_{TM} - VMA} \right) )</td>
</tr>
<tr>
<td>Where:</td>
</tr>
<tr>
<td>( \eta(4.75 – 2.36) )</td>
</tr>
<tr>
<td>( PP_{2.36} )</td>
</tr>
<tr>
<td>( PP_{4.75} )</td>
</tr>
<tr>
<td>VMA</td>
</tr>
<tr>
<td>( V_{TM} )</td>
</tr>
</tbody>
</table>
correlating them with permeability or inferred permeability.

**Revised Bailey Method Ratios (BMRs) based on contiguous aggregate fractions**

As described in the preceding section, it is possible to calculate porosity for contiguous aggregate fractions using Equation 7 (Denneman et al. 2007). The typical aggregate sieve fractions used in South Africa, as well as the new South African National Standards (SANS) sieve sizes, basically tend to have a size ratio (fine/coarse preceding) of 0.5 particularly at the meso and micro nesting levels. However, at the macro level and a portion of the midi or rather meso level the ratio of consecutive sieve sizes is closer to 0.8, implying there are more sieve descriptions at these macro and midi or rather meso skeleton levels. Typically sieve sizes at 3.35 mm, 1.18 mm, 0.85 mm, 0.71 mm, 0.425 mm, 0.335 mm, 0.25 mm, 0.121 mm, 0.125 mm, 0.106 mm, 0.09 mm, 0.63 mm, 0.045 mm, 0.038 mm, 0.032 mm and 0.020 mm exist within the ASTM sieve ranges and are absent in the sieve sizes used. As may be shown later, the use of these sieve sizes can greatly enhance the analysis at the micro level of the whole grading envelope.

By exploring the consecutive contiguous aggregate fractions as per the SANS sieve openings an expanded set of BMRs could be described. These are referred to as rational BMRs and are shown in Table 2 for the aggregate grading matrix at macro, midi or meso and micro levels. It shows the rational BMRs in the first column and the inverse of the rational BMRs in the second column. The logic of this latter inverse rational BMRs will be discussed later.

As discussed above Al Mosawe et al. (2015) already demonstrated the potential for these rational BMRs for structural strength indicators. It has also been demonstrated by Horak et al. (2017b) on reworked published data sets. Komba et al. (2019) recently used these rational BMRs with success to provide better indicators for HMA compactability efficiency. Only limited work has been done on the full set of the rational BMRs linkage with permeability by Horak et al. (2017a). Machine

<table>
<thead>
<tr>
<th><strong>Table 2 Rational and revised Bailey ratios with good correlation with DASR porosity parameters</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matrix level</strong></td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Macro</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Midi-Meso</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Micro</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 2 Rational and revised Bailey ratios with good correlation with DASR porosity parameters

<table>
<thead>
<tr>
<th><strong>Matrix level</strong></th>
<th><strong>Rational BMRs</strong></th>
<th><strong>Inverse Rational BMRs to align with BAP principles</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro</td>
<td>PN = (%NMPS – %HS) / (%NMPS – %HS)</td>
<td>PO = (%100 – %NMPS) / (%NMPS – %HS)</td>
</tr>
<tr>
<td></td>
<td>PO = (%100 – %NMPS) / (%NMPS – %NMPS)</td>
<td>%Retained on NMPS = %Pluggers (PN)</td>
</tr>
<tr>
<td></td>
<td>CA = (%HS – %PCS) / (%100 – %HS)</td>
<td>CA = %Retained on NMPS and HS / %Interceptors</td>
</tr>
<tr>
<td></td>
<td>%Interceptors / %All pluggers</td>
<td>%All Pluggers (P(N) + P(O)) / %Retained on PCS</td>
</tr>
<tr>
<td>Midi-Meso</td>
<td>Cf = (%PCS – %SCS) / (%HS – %PCS)</td>
<td>Fc = (%HS – %PCS) / (%PCS – %SCS)</td>
</tr>
<tr>
<td></td>
<td>Fc = (%HS – %PCS) / (%PCS – %SCS)</td>
<td>%Coarse portion of fines / %Interceptors</td>
</tr>
<tr>
<td></td>
<td>F = %PCS / (%NMPS – %PCS)</td>
<td>C = (%NMPS – %PCS) / (%Plugger(PN) + Interceptors)</td>
</tr>
<tr>
<td></td>
<td>I = %HS – %PCS / (%NMPS – %HS)</td>
<td>%Interceptors / %Pluggers (PN)</td>
</tr>
<tr>
<td></td>
<td>%Interceptors / %Pluggers (PN)</td>
<td>%Retained on HS / %Pluggers</td>
</tr>
<tr>
<td>Micro</td>
<td>FAcm = (%SCS – %TCS) / (%PCS – %SCS)</td>
<td>FArcm = (%PCS – %SCS) / (%SCS – %TCS)</td>
</tr>
<tr>
<td></td>
<td>FAcm = (%SCS – %TCS) / (%PCS – %SCS)</td>
<td>%Medium fine of fines / %Coarse of fines</td>
</tr>
<tr>
<td></td>
<td>FAcm = (%STCS – %Filler) / (%SCS – %STCS)</td>
<td>FAcm = (%SCS – %TCS) / (%TCS – %Filler)</td>
</tr>
<tr>
<td></td>
<td>FAcm = (%STCS – %Filler) / (%SCS – %STCS)</td>
<td>%Medium fine of fines / %Coarse of fines</td>
</tr>
<tr>
<td></td>
<td>FAcm = (%STCS – %Filler) / (%SCS – %STCS)</td>
<td>%Fine of fines / %Medium fine of fines</td>
</tr>
</tbody>
</table>

Table 2 Rational and revised Bailey ratios with good correlation with DASR porosity parameters
learning techniques have recently been used to develop permeability prediction of HMA based on the principles and approach described in this paper (Blaauw et al. 2019), and Cromhout (2018) used it with success on QA/QC large data set analyses. The logic of the approach developed will be described next by focusing on identifying parameters linked to interconnected voids as main indicator of permeability potential.

**IMPROVED INTERCONNECTED VOID MONITORING WITH BINARY AGGREGATE PACKING (BAP) PRINCIPLES**

**BAP origin and conversion to asphalt mix design**
The original work by Furnas (1928) with binary and ternary aggregate fraction combinations was based on experimental work and observations of beds of broken coarse and fine solids (aggregates which form the basis of the understanding of Binary Aggregate Packing (BAP) principles). His work was developed further by various researchers in pursuit of efficient aggregate packing of concrete mixes, and was subsequently transferred to asphalt mix design by a number of researchers (Baron & Sauterey 1982; Francken & Vanelstraete 1993; De Larrard & Sedran 1994; Olard & Perraton 2010; Butcher & Van Loon 2013). Furnas originally determined the basic volumetric particle diameter ratio of 0.20 to 0.23 of fine aggregate to coarse aggregate size ratio below which the voids between the coarse aggregate particles can be filled by the fine particles at optimum packing density and obviously optimum structural strength. This is the same principle (0.22 ratio fine/large) that was subsequently used by the Bailey method as basis (Vavrik et al. 2001). Low porosity is achieved at or below this size ratio and thus the implied structural strength provided via the aggregate skeleton is strong.

In the case of the DASR method the volumetric arrangement of aggregate fractions is the combination of a range of contiguous aggregate fractions (mostly more than two) which have the lowest or optimum porosity (preferably below 0.5 and typically closer to 0.4). This efficient aggregate packing is obviously reliant on the basics of the binary aggregate packing principles. Olard and Perraton (2010), and Butcher and van Loon (2013) describe how this theoretical basis of BAP, expressed in terms of size and volume ratios, can be used to achieve optimum packing for asphalt mix design at the lowest porosity.

**The BAP porosity triangle**
The real insight provided by this approach regarding the aggregate interaction at each skeleton subset level comes from the use of BAP porosity ‘triangle’ diagrams (Francken & Vanelstraete 1993; Olard & Perraton 2010; Shivaprasad et al. 2014; Butcher & Van Loon 2013; De Larrard & Sedran 1999; Mangulkar et al. 2013; Perraton et al. 2007). These BAP porosity triangle diagrams were developed to relate porosity or void indices or specific voids to various proportions of coarse aggregates and fine aggregates. The basic BAP triangle diagram is illustrated in Figure 1 and is based on the original work by Furnas (1928). The concepts of voids, porosity and specific voids are interrelated, and in this paper are assumed to entail the same concept of an index of voids between solids prior to filling with binder and in general is referred to as porosity (Butcher & Van Loon 2013).

The porosity of the fine aggregate (left Y-axis in Figure 1) and the porosity of the coarse aggregate (right side Y-axis in Figure 1) are higher than the porosity of the combination of the two fractions as a binary fraction combination. Olard (2015) demonstrated that these size ratios (fine/coarse) form a family of curves (concave functions) with variable combined porosities lower than the porosity of either fraction separately. The porosities of the combined binary fractions also vary in terms of the proportion of the volumetric ratio of coarse/fine (X-axis in Figure 1) along these concave form functions. In Figure 1 it is shown that the optimum or lowest point of porosity achievable, the dilation point, is found for any size ratio (fine/coarse) concave function proportion; at a specific coarse-to-fine aggregate volume ratio the two aggregate sizes combine (see the X-axis).

The concave curves tend to be much more pointed (lower porosity) below the 0.2 value (equivalent to Bailey ratio of 0.22) for the fine/coarse diameter ratio range, therefore the densest possible packing. The reverse is also true, that if the fine/coarse size ratio exceeds 0.2, the concave function flattens out and the proportion of coarse aggregate per volume (or mass) becomes increasingly less influential with regard to reduction in the combined fraction porosity as this ratio approaches 0.5. Thus, the logic and basis of the Bailey control sieves are confirmed to correlate with the binary packing triangles as well. The size ratio determines the shape of the concave function and shows that, for the flatter shaped concave functions, the potential to lower porosity is reduced irrespective of volume ratio changes on the X-axis. Most of the contiguous aggregate fractions along the

![Figure 1 Typical binary aggregate combination porosity influence with varying diameter ratios and proportion of coarse aggregate (Olard et al. 2010, based on Furnas 1928)'](image-url)
entire grading envelope of a typical continuously graded mix used in South Africa (see Table 2) are spaced at the 0.5 factor or more, which implies that the upper flat concave shaped function is mostly applicable. This implies that there may be limited scope in lowering porosity with the known sieve sizes at the midi or rather meso and micro levels of the grading envelope.

**Wall and loosening effects on porosity in Binary Aggregate Packing (BAP) triangles**

Under ideal conditions the two dotted lines in Figure 1 are the outer boundaries of the possible physical zone of potential porosity achievement between the fine and coarse aggregate in the binary combination. The basic concave form of curves was developed by Baron and Sauterey (1982) based on the original experimental work of Furnas (1928) and the influence in practice on porosity due to a retardation effect on porosity reduction. This resistance to porosity reduction was found to be due to two physical phenomena: (1) *The wall or boundary effect*, and (2) *The loosening or disruptor effect*. Both concepts are illustrated in Figure 2 (Knop & Peled 2016) for a realistic aggregate mix situation.

*The wall effect* is due to the void increase caused by the boundary or wall created by the large aggregate (1 in Figure 2) acting as boundary to the finer sized aggregates (2 in Figure 2), and the voids against the large aggregate “wall” being prevented from being infilled by other medium-sized or fine particles. This “wall effect” is creating additional unfilled void space next to the larger aggregate boundary, and in effect resisting the lowering of porosity. It is also obvious that these voids against the “wall” have a higher probability to be interconnected along a defined plane (the wall) with obvious influence on water flow and permeability.

*The loosening or disruptor effect* is due to the increase of the fine or smaller aggregate particles in the large or medium aggregate voids to the point where they “overfill the voids” and in effect start separating or pushing the larger aggregates apart. In effect, the void index or porosity is again prevented from being lowered. Therefore, these additional voids created would also increase the probability of being interconnected on their own or enhance the interconnectedness, particularly if linking up with the wall effect and creating “additional interconnected voids”.

These phenomena hold the key to monitoring the interconnectedness of voids. The two effects combined for a specific diameter ratio concave function are shown in the coarse/fine volume ratio variance in the basic binary triangle diagram in Figure 3 (Olard 2015). The zones where the porosity is restricted from decreasing
are clearly indicated. The relations in the binary packing triangle diagrams are shifted upwards from the theoretical maximum linear relationships to a concave form porosity function as illustrated in Figure 3. The zones where the wall effect and the loosening effect lead to increased porosity versus the minimum potential with the theoretical function are clearly identified by the hatched areas. The most obvious differentiator is the varying coarse/fine aggregate proportion by volume in this binary fraction combination as it varies on the X-axis. The zone around the dilation point position on the X-axis is shown in Figure 3. This dilation zone clearly leads to a maximisation of the potential interconnectedness of these additional voids created in the aggregate matrix. In this zone from left to right on the X-axis, and leading up to the dilation point, the wall effect increasingly dominates (0 to 0.6 coarse in the mix). In the dilation point zone (0.6 to 0.8 of coarse in the mix) the “additional voids” with a higher probability of being interconnected are influenced by both the wall effects and loosening effects increasing the combined porosity or resistance to porosity reduction. It is noted that the combined porosity rate of increase is the highest at the dilation point (0.8). To the right side of the dilation point the percentage of coarse aggregate dominates in the mix (right hand side in Figure 3) and where the loosening effect contribution tends to reduce while the wall effect continues to increase and the combined “additional porosity” tends to reduce.

**Direct influence of BAP ratios on permeability**

Experimental work by Mota et al (2014) on binary aggregate combinations for filter beds is illustrated in Figure 4. It illustrates how Mota et al showed the coefficient of permeability (Y-axis in Figure 4(a)) is affected by the coarse/fine aggregate volume ratios (X-axis in Figure 4(a)), as well as the diameter ratios. In Figure 4(a) the trend shown illustrates that coarse/fine volume proportion ratios from 0 to 0.6 are virtually insensitive to the permeability measured. This is the zone where the fine aggregate volume fraction will be proportionally higher in the mixes as per Figure 3.

Figure 4(a) was determined with a fixed diameter ratio (coarse/fine ratio) of 0.22. This translates to the inverse as per Figure 1 of fine/coarse of 0.1, which implies low porosity. For binary mixes with coarse/fine volume ratios above 0.6 (therefore more coarse material in the mix) the permeability coefficient starts to increase significantly. Above the ratio of 0.8 the increasing trend is exponential. Thus, coarse/fine aggregate volume ratios of 0.6 and 0.8 can be viewed as threshold values. This threshold value also holds true even if the binary fraction size or diameter ratio is varied as illustrated in Figure 4(b). As was illustrated in Figure 2, the dilation point or optimum zone of the concave curve (coarse/fine ratio of 0.6 to 0.8) is levelled out if the D/d size ratios decrease, and obviously if the inverse d/D ratios increase as illustrated in Figure 2.

This basic observation of coarse aggregate size influence on porosity and permeability has already been demonstrated by Cooley et al (2002) via permeability testing on HMA mixes. They found that HMA mixes with NMPS in the range of 19 mm and 25 mm show increased permeability at low void contents (even 4% voids as threshold value) for asphalt mixes. The norm for HMA with 12.5 mm and lower NMPS in the mix is a threshold value of 7% voids in the mix (Cooley et al 2002), where after permeability measured will increase exponentially with further increase in air void content. This effect of NMPS in the HMA clearly demonstrates the wall effect (literally against the large aggregate surface) on increasing the interconnected voids in the mix.

It appears that an increase in permeability (see Figure 4(a)) may be linked to the wall effect and the loosening effect for coarser mixes when the threshold value of 0.6 coarse proportion in the mix is exceeded. This tends to confirm the 0.6 as lower trigger value described before. It appears that the size ratio (d/D) has a significant effect on permeability, but the form of the curves with volumetric ratio on the X-axis shows classical “lift off” behaviour around the 0.8 value previously described to be the dilation point in Figures 1 and 3. Therefore the 0.6 and 0.8 values of the volumetric ratios can be used as possible threshold values for permeability control at the various levels of macro, midi or rather meso and micro subset infill skeletons.

Several researchers state: “... when the specific gravity of the material is the same for all the aggregate fractions, the
volumetric ratio or proportion of coarse and fine aggregates can be taken as equivalent to the proportion of percentages retained on the coarse and the fine aggregate sieves..." as the basis of this simplified calculation (Butcher & Van Loon 2013).

Therefore, it should be possible to benchmark the various binary aggregate mass combinations making up the macro, meso and micro levels to monitor where adjustments can be made to aggregate gradings to achieve packing efficiency, as well as provide permeability control in a rational fashion.

**Proposed mass ratio range as benchmark for interconnectedness of voids**

The discussion on the wall effect, loosening effect on porosity, and permeability indicates that the BAP “triangle diagram” (Figures 1 and 3) can be used as a benchmark measurement tool by combining the trends also shown in Figure 4. The trend in porosity versus the horizontal axis of the BAP triangle can be simplified as illustrated in Figure 5. Interconnectedness of voids, due to the wall and loosening effects described before, is also shown regarding their zones of dominance for these “additional and potentially interconnected voids”. This conceptual trend in porosity is further enhanced by superimposing the implied RAG benchmark criteria for additional voids due to retardation, implying void interconnectedness. Red is high probability (BAP coarse aggregate mass portion is 0.6 to 0.8), Amber is medium probability (BAP mass coarse portion is 0.4 to 0.6) and Green is low probability (BAP mass coarse portion is 0 to 0.4). The benchmark ranges of the coarse aggregate percentage (horizontal scale), moving from left to right, first show the wall effect alone and then the addition of the loosening effect (increasingly combined). This helps to facilitate the evaluation of BAP ratios in terms of porosity or relative increase in interconnectedness, and therefore indirectly the permeability potential.

In order to convert the percentage of coarse in the BAP combination on the horizontal scale of Figure 5 it needs to be converted to coarse/fine as a mass ratio. In Figure 6 the relationship is shown graphically to illustrate how this conversion is achieved. This is to ensure that the porosity and BAP coarse/fine mass ratios (inverse of rational BMRs defined in Table 3) can be set on the same scale for direct comparisons.

**TESTING THE BAP BENCHMARK CRITERIA WITH REWORKED PUBLISHED DATA SETS**

**Description of the reworked published data sets**

Three data sets were initially available for analysis to demonstrate the effect of the coarse/fine ratios and the associated porosities on permeability. The three data sets are from Denneman et al (2007), Al Mosawe et al (2015), and Sadasivam and Khosa (2006), previously also reworked by Horak et al (2017a&b).

Horak et al (2017a&b) previously found that trends are best observed by focusing on the high and low permeability extremes of the data set, as the middle permeability ranges gave confusing results due to the now known randomness or variance in void interconnectedness. It needs to be stressed that these various data sets do not always have actual permeability measurements to enable a clear discernment between low permeability and high permeability data.
sets. Typically, Denneman et al. (2007) only expanded on criteria developed by Sadasivam and Khosa (2006). Thus, the Denneman et al. (2007) data sets related to probable permeability values indirectly and made a distinction between asphalt mixes with good and those with poor performance (primarily based on cracking observations), while rutting resistance was good in both cases. Sadasivam and Khosa (2006) did have actual permeability measurements, which allowed a high permeability data set and low or impermeable data set to be made up from the original.

Al Mosawe et al. (2015) showed void content (%) in their published data set. It is known that high void content correlates well with higher permeability, typically above 6% (e.g. air voids % above 6%) for an NMPS of 12.5 mm where interconnectedness of voids starts and therefore permeability starts and increases (Cooley et al. 2002). In the analysis and reworking of data a distinction could be made between probable high permeability and medium to low permeability based on low void content and high void content. Therefore, only these upper and lower extremes of voids or permeability were used in the reworked data. The analysis was further simplified here by using only the average values for these categories to simplify the demonstration.

**Example of macro to micro level binary fractions**

*Explaining the benchmark comparison graphs*

The data sets presented by Denneman et al. (2007), which have been reworked before to determine the Bailey ratios and DASR porosities of ranges of aggregate fractions (Horak et al. 2017a), were used as worst-case example of data quality with regard to permeability. The data set was reworked as explained above with rational inverse BMRs to represent the binary aggregate fraction contiguous combinations for the whole spread of macro, midi or rather meso and micro nested aggregate subsets. The various aggregate fraction combination porosities are calculated in terms of coarse/fine mass ratios by means of Equation 7. Figures 7 and 8 use a single ratio axis covering the porosity (Y-axis) of the triangle binary packing diagrams, as well as the inverse rational BMR (which is the ratio of coarse/fine by mass (% retained per fraction in the ratio) measured against the conversion as illustrated in Figure 6. The third bar indicator shown is the product of the mass ratio and porosity as a further possible combination indicator of the two calculated factors.

The RAG zones, as defined in Figures 5 and 6, are superimposed on these graphs to help identify binary fraction mass ratios which may extend into the Amber zone (between 0.6 and approaching 1.5 mass ratio) where the wall effect dominates, or the Red zone (above the mass ratio of 1.5) where both the wall effect and the loosening effect are at play here. Therefore, this should be interpreted with the next level inverse rational BMR, FArcm, which shows that the potential voids caused by the wall and loosening effects in Fc/Cf have been filled, as FArcm has a mass ratio on the border between Green and Amber, therefore limited possibility of interconnected voids and low permeability probability.

*Poor-performing mix*

The Denneman et al. (2007) dataset of poor-performing asphalt mixes is shown in Figure 8. The coarse/fine retained sieve mass ratios, the associated combined porosity values and the product of these two factors are shown. It is the consecutive contiguous coarse/fine mass ratios of the parameters PN/I, Fc/Cf and FArcm.
that have values in the Amber zone. This implies that the macro level PO/PN at the plugger (oversize and normal size) zone of the aggregate grading envelope has no additional voids created, but the void creation due to the wall and loosening effects start at the midi or rather meso level with the PN/I which is not filled by $F_c/C_f$ and in turn finally not by $F_{Arcm}$. Therefore this dataset shows there is a high probability of interconnected voids not being filled at the micro level, and therefore a high probability of it being permeable. Bhattacharjee and Mallick (2002), in their search for a better “handle” on permeability, focused on the finer portion of the grading, as well and stated: “Porosity is significantly affected by percent of material passing 2.36mm sieve.”

This tends to confirm the Bailey mantra that “the voids are in the fines” where interconnectedness finally happens and therefore the probability of permeability is increased. This observation can be paraphrased as: “The void infill back stops at the bottom of the grading envelope and if not filled it will create opportunity for interconnected voids.”

To demonstrate the statement about the micro level as the most probable level where voids may not be effectively filled, the micro level inverse rational BMRs ($F_{Arcm}$) for all three data sets were calculated for the good-performing as well as poor-performing extremes of their data sets, and are shown in Figure 9.

The Denneman et al (2007) data sets show that the porosity values are not significantly different if the good- and poor-performing data sets are compared. The porosity range is between 0.6 and 0.7, implying that this binary aggregate combination may have adequate density achievement at this level. However, the mass ratio (coarse/fine) for $F_{Arcm}$ clearly can discern between the poor- and good-performing datasets, as the values differ significantly. The bad-performing dataset $F_{Arcm}$ mass ratio value shows a strong wall effect impact on porosity and is well within the Amber zone. The product value ($Product = F_{Arcm} \times Porosity$) obviously shows the dominance of the $F_{Arcm}$ and implies that the potential for interconnected voids may be high, thus increasing the potential for water flow or higher permeability values.

The Sadasivam and Khosa (2006) reworked data shows the same trend or sensitivity to coarse/fine mass ratio and combined porosity values at the micro nested aggregate skeleton subset level. In this case permeability was measured, which indicates that the $F_{Arcm}$ mass ratio may in fact be the real discriminating factor between permeable and low permeable mixes. The associated porosity values are in the same range for the high and low permeability values, therefore confirming porosity alone is not a reliable discerning factor of permeability potential. The mass ratio of $F_{Arcm}$ of the permeable dataset is clearly exposed to both the wall and loosening effect, being in the Amber zone close to the Red zone. This implies a higher possibility of interconnected voids. The low permeability $F_{Arcm}$ mass ratio is at the border between Green and Amber, implying low interconnected voids. The product value ($Product = F_{Arcm} \times Porosity$) confirms porosity exposed to the wall, as well as that the loosening effect may be indicating a tendency for the voids in the porosity to have interconnectedness of voids.

Figure 8 Poor performing dataset reworked porosity and coarse to fine proportion analysis of micro-level of all reworked data sets

![Figure 8](image)

Figure 9 SCS to TCS fraction ratio and porosity and product

![Figure 9](image)
The Bailey Method Ratios (BMRs) were reconfigured by using the void content in the mix are less precise than desired. The reason is that the voids represented by the medium-range porosity are in fact exposed to both the wall and the loosening effects. This dataset may therefore have a higher probability of interconnected voids, and as a result a higher permeability probability. The product in this case is not necessarily a good indicator of permeability, due to the confusing result of the porosity. Therefore the product as an indicator is not viewed as a reliable indicator of interconnectedness of voids and should not be used on its own.

In this respect Bhattacharjee and Mallick (2002) support the use of porosity on its own as a better indicator of permeability potential, but in their case it implies the porosity of the whole gradation.

CONCLUSIONS

Voids alone cannot effectively measure permeability potential

Permeability as durability indicator can be measured under controlled conditions in the laboratory, but is more difficult to measure in the field for quality control and assurance purposes. Indirect measures using the void content in the mix are less precise than desired. The reason is that the voids may not be continuously interconnected, and for the same void content permeability can be either non-existent or already even highly probable. Normal density specifications also imply that the range of voids specified may stretch beyond limits, which tend to show an exponential increase in permeability once the voids become increasingly interconnected. Therefore additional indicators linked with the aggregate grading need to be identified to help monitor permeability potential.

Reconfiguring the Bailey Method Ratios to adhere to contiguous aggregate fractions

The Bailey Method Ratios (BMRs) were investigated as potential indicators of interconnectedness and permeability with limited clear linkages or correlation. In studying the Dominant Aggregate Size Range (DASR) it was observed that contiguous aggregate fractions can be used to calculate porosity as an indication of density. The BMRs were therefore reconfigured by using only contiguous fractions in the calculation of a whole expanded set of revised or rational BMRs. Porosity thus calculated for rational BMRs also runs into the same issue as with voids in the mix. Porosity alone is also not an indication of permeability, but rather the interconnectedness of the voids. Spatial composition is defined (SABITA 1999) as “... the three-dimensional modelling of asphalt at macro, meso and micro level. Spatial composition is the only volumetric way to obtain a holistic understanding of the complex interactions between the several components, being bitumen, coarse and fine aggregate, filler, air and water.” Therefore the reconfigured and expanded BMRs’ biggest contribution is the provision of a better definition of structural strength, as well as possibly permeability. Future work should include the aspects of effective film thickness and mastic in the aggregate spatial composition.

Linkage with the Binary Aggregate Packing concept

The “deconstruction” or nesting of the total aggregate matrix in sub-skeletons at the macro, meso and micro levels of the total grading curve of HMA provides insight into the actual packing of the voids at each aggregate fraction level. The Binary Aggregate Packing (BAP) concept was investigated and found to describe the actual packing of the aggregate fractions and resultant voids at the various nested subset levels quite well.

The BAP concept is described in terms of the coarse/fine retained mass ratio, characteristic concave functions based on size ratio, porosity of individual fractions and porosity of the combined aggregate fractions in the BAP porosity triangle diagrams. The curves of the coarse/fine combination versus porosity have characteristic concave shapes, influenced by the ratio of the fine/coarse aggregate size ratios.

It was observed that the rational BMRs are the inverse of the rational BPA mass ratios. These inverse rational BMRs allow the description of the void packing or infill efficiencies at the macro, midi or rather meso and micro nested subset levels of a grading envelope. It therefore allows monitoring of structural, as well as porosity and permeability, control aspects.

Phenomena in BAP porosity triangles that can be linked to void interconnectedness

The two phenomena that restrict porosity reduction (therefore also a denser mix) in BAP aggregate fractions studied in the BAP triangle of porosity determination were identified to give a better indication of potential for interconnectedness of voids. It was shown how the coarse/fine mass ratio has a dilation zone, whereas the wall effect and the loosening /dislodging effects tend to restrict porosity and have a tendency to be interconnected. These effects were transferred from these BAP triangle diagrams to allow a typical colour-coded three-tiered RAG benchmark indication for void interconnectivity by using the coarse/ fine mass ratios with porosity for various contiguous inverse rational BMRs.

Demonstration of mass ratio benchmark potential

The potential of the benchmark criteria was demonstrated with the reworking of published data sets with directly or indirectly measured permeability. Only the extremes of the data sets – high permeability or low permeability – were used, due to the statistical variance of voids in the mix in the midrange described before. The analyses showed that the coarse/fine mass inverse rational BMRs function independently from porosity values and tend to correlate well with void interconnectedness and therefore permeability potential. It was also found that the Bailey method mantra (that the voids are in the fines) holds the key, and that successive levels of void infill down to the micro level may be the best indicator of permeability, due to the mass ratio at the lowest level still not properly infilled with a graded finer aggregate. The critique on the current SANS sieve system is the lack of in-between sieve sizes at the midi or rather meso and micro levels.
This may imply that structural design overemphasis leads to a lack of control and analysis opportunity at the these lower levels, which may negatively influence permeability control and awareness.

**Simple spreadsheet calculations of the inverse of rational BMRs can be used**

It is clear that a simple spreadsheet analysis of the inverse of rational BMRs derived from contiguous aggregate fractions on the grading envelope can be used to assist and monitor the potential for permeability. The link of rational BMRs with porosity to the grading envelope can be used to assist and analyze the inverse of rational BMRs derived whether a mix may be prone to permeability, even though it may meet rut resistance criteria.

**Additional testing**

Even though only average values were used in the demonstration of the concepts with reworked dataset extremes, the use of the inverse rational BMRs with full datasets using Probability Density Functions (PDFs) (Cromhout 2018) has proved itself as a viable benchmark that can discern between probable permeable and non-permeable using as-built and QA/QC data. Other indicators of probable permeability indications are also available to be calculated on a continuous basis also linked to normal QA/QC or as-built and grading envelope information (not discussed here). These approaches can be used in a typical forensic investigation mode (NCHRP 2013) to identify areas of constructed asphalt surface that benchmark or predict as marginal to strongly permeable. Such areas of asphalt pavement can be isolated or demarcated to be investigated in more detail. This zooming in on more detailed field surveys (Horak et al. 2015) may typically include:

- Normal coring with density and void determination
- Marvil field permeability testing
- Laboratory permeability testing
- Visual evaluation of the core surface void interconnectedness appearance
- Computed thermography scans and analysis (Horak et al. 2015).

**REFERENCES**


Blaauw, S A, Maina, J W & Horak, E 2019. Towards a hot mix asphalt design model for the prediction of durability. Submitted for publication to *Construction and Building Materials*.


Validating traffic models using large-scale Automatic Number Plate Recognition (ANPR) data

A Robinson, C Venter

The development of reliable strategic traffic models relies on comprehensive and accurate data, but traditional survey methods are time-consuming and expensive. Manual surveys often yield small samples that require estimated expansion factors to enable the data to represent the population. Modellers have turned to new data sourced from various electronic devices to improve the reliability of the data. Automatic Number Plate Recognition (ANPR) data is one such data source that can be used to extract travel time, speed and partial origin-destination (OD) information. This study assesses ANPR data in terms of its comprehensiveness and accuracy, and shows how it can be used for the validation of strategic traffic models. Data was obtained from the Gauteng freeway system’s Open Road Tolling (ORT) gantries for a period of several months. A new methodology is developed to process traffic model outputs such that they are directly comparable to the partial origin-destination outputs derived from the ANPR data. It is shown that comparing the model distribution against observed ANPR data highlights potential trip distribution issues that are not detected using standard model validation techniques.

INTRODUCTION

Automatic Number Plate Recognition (ANPR) entails the automated recording of the number plate, date/time and location of each vehicle that passes a roadside camera, using vehicle number plate recognition software. Records of individual vehicles that pass multiple cameras can be matched to determine the path of the vehicle and calculate travel times between the survey locations. If cameras are in a closed cordon, the origin and destination of external trips passing through the cordon can be determined. A series of ANPR cameras along a route, or at strategic locations throughout a network, would not observe every vehicle upon entry and exit to the network, and constitutes an open format number plate survey. Both closed and open format ANPR data have the potential to provide information that can be useful during the development of strategic traffic models, in ways that are not possible with other sources of traffic data. Comprehensive traffic observations from loop detectors, like ANPR, provide link speed and volume information which is useful during the calibration and validation of traffic models. But the additional ability of ANPR to track individual vehicles from point to point also provides potentially useful data on the distribution of trips through the network. While this constitutes partial rather than comprehensive origin-destination (OD) data, it may still serve as an additional independent data set against which model outputs can be validated. ANPR data has rarely been used in this way.

The objective of this paper is to examine the use of ANPR data for traffic model validation in terms of its comprehensiveness and accuracy. ANPR data is provided by the South African National Roads Agency SOC Ltd (SANRAL) from the Open Road Tolling (ORT) system deployed on the Gauteng Freeway Improvement Project (GFIP). Selected link volumes and journey times are, for demonstration purposes, compared with the GFIP traffic model’s 2015 forecasts. In addition, the trip distribution characteristics of the ANPR data are exploited by extracting partial OD and trip length distribution metrics for comparison with modelled quantities. This required the development of a new methodology to process traffic model outputs such that they are directly comparable to ANPR-derived partial OD data. This is a feature of model validation that has not been found in previous studies.

Keywords: ANPR, traffic models, origin-destination, validation
The paper first provides a brief overview of the literature regarding techniques for developing and validating strategic traffic models, and specifically matrix estimation. It then describes the study context in terms of the GFIP, and the extent and accuracy of the ANPR data that is collected. The development of a technique for extracting suitable data from strategic models for comparison against ANPR data is presented, and implemented using the GFIP data to reach conclusions regarding the validity of the GFIP model forecasts. Finally, drawing on this work, the potential strengths and pitfalls of using ANPR data for traffic model improvement are discussed.

DATA REQUIREMENTS FOR STRATEGIC TRAFFIC MODEL DEVELOPMENT

Traditional strategic traffic models are developed using the standard four-step process. For this, these models typically require the following data sets:

- Road network geometric information to develop the core network
- Land use data to determine the trip generation
- Origin-destination (OD) data to derive the trip distribution
- Generalised cost data to determine mode and route choice
- Speed and traffic count data for the volume-delay relationships used in assignments
- Journey time and traffic counts to calibrate and validate the model.

OD trip matrices are fundamental inputs into traffic studies and traffic models. As observed data only provides information to form partial matrices, the development or synthesis of full trip matrices has been the focus of many studies dating back to John Wootton in 1972 (Kirby 1979). Data is obtained through household, roadside or other interview survey techniques. A full “prior” trip matrix is then determined using distribution functions derived from the survey data, and estimated and calibrated from using other observed data such as traffic counts. Model validation must be undertaken using independent data not used in the model development and calibration. The validation of the trip distribution is based on the comparison of partial OD volumes (after calibration) with cordon and screen-line counts, and of modelled trip length frequency distributions (TLFD) with those observed in surveys and previous studies.

Given the difficulties of estimating base year OD matrices from incomplete information, some researchers (Willumsen 1981; Fisk 1989; Tamin & Willumsen 1989) have sought to maximise the use of additional information such as traffic counts to produce cost-effective trip matrix estimations. The problem remains that the number of independent traffic counts are typically insufficient to produce a unique OD matrix. To create a unique matrix with N zones one requires N² fully balanced traffic counts, all taken at the same time with no other sinks and sources other than the zone connectors (Ortuzar & Willumsen 1998). This is an impossible task in large-scale models.

This is where Automated Number Plate Recognition holds promise, as it is possible to generate larger sets of data for use during matrix estimation, distribution function calibration, and validation. In a simple form, the concept of using ANPR data in OD matrix estimation is described by Ramirez et al (2013) where it was applied in a limited way at localised intersections. Castillo et al (2008), Minguez et al (2010), and Hadavi and Shafahi (2016) researched the optimisation of camera locations to maximise the potential coverage and usefulness of the data obtained. Asakura et al (2000), Dixon and Rilett (2005), and Van Vuren and Carey (2011) used ANPR to analyse trips on motorways to derive through-trips and interchange-to-interchange trips. They also derived methods to expand samples where the cameras did not cover all lanes. Sun et al (2014) developed metrics for tracing vehicles passing cameras while travelling on a city network, and Himayounfar et al (2011) assessed travel patterns to benchmark normal behaviour to highlight suspicious drivers for law enforcement.

Cameras with ANPR capability

Volumetric vehicle classification systems.

As each vehicle passes under a toll gantry, the vehicle’s number plate, its toll classification (SANRAL 2018), date/time stamp and gantry number are recorded. SANRAL provided the ANPR data used in this research.

The GFIP strategic traffic and toll revenue forecasting model was developed in 2007 to determine the impact of the freeway upgrades and tolling on freeway traffic volumes and the surrounding road network.

The traffic model was developed using the SATURN (Van Vliet 2015) traffic modelling software and used the provincial GTS200 (Gauteng Department of Roads and Public Works 2006) traffic model as a starting point. The model was updated and calibrated to 2006 base year traffic data including:

- Journey time surveys from the freeways and major competing routes
- Land use data, interpolated between the 2001 census data and the 2010 land use forecasts
- Revised trip generation rates
- Revised average trip lengths for light and heavy vehicles
- Approximately 600 classified traffic counts from 2006.

The forecast years were 2010, 2015 and 2025.

THE GAUTENG FREEWAY IMPROVEMENT PROJECT (GFIP)

The GFIP comprised the upgrading and tolling of 201 km of urban freeways in Gauteng, South Africa, and included the addition of carriageway lanes and the upgrading of interchanges. The freeways are tolled using an Open Road Tolling (ORT) system incorporating 42 directional toll gantries at approximately 10 km spacing. The GFIP freeway network and the locations of the toll gantries on the freeway network are depicted in Figure 1.

Equipment on the toll gantries includes the following systems required for toll collection:

- Cameras with ANPR capability
- Volumetric vehicle classification systems.

No examples have been found in the literature of the use of ANPR data collected over a large area – for the validation of traffic models. The ANPR data produced by the Open Road Tolling system in Gauteng provides an opportunity for testing the feasibility and usefulness of such an application of what is essentially by-product or “exhaust” data from the tolling infrastructure.
Figure 1 GFIP network and ORT gantry locations (SANRAL 2018)
**REVIEW OF THE ANPR DATA**

### Extent of the ANPR data

Monthly ANPR data was provided in text files. Prior to receiving the data, the vehicle licence number (VLN) was replaced with a random number VLN ID to anonymise the data to comply with the Protection of Personal Information Act, 2013. Each vehicle’s VLN ID remained constant within each month’s data to ensure that vehicles could be tracked through the network over consecutive days.

Table 1 provides the total number of gantry entries per month between February 2014 and July 2015. Approximately 71 million ANPR records per month were obtained from all 42 gantry locations over this period.

### ANPR data accuracy

The accuracy of the ANPR data was assessed in two ways. Firstly, the data was compared to equivalent electronic traffic counts obtained from permanent counting stations located along the freeway network, and secondly, based on an interrogation of the completeness of the data in terms of the ability to track vehicles through the network which would be affected by unreliable number plate records.

SANRAL has installed electronic traffic counters at freeway interchanges as part of its Comprehensive Traffic Observation (CTO) programme. The counters at the interchanges upstream of each toll gantry were used and compared to the gantry’s ANPR data. The equivalent average hourly weekday and weekend traffic counts were extracted from each database for each gantry location and compared. The average hourly volumes were calculated by adding the hourly volumes for every weekday or weekend day and dividing by the number of weekdays and weekend days in the month. Any missing data was recorded as zero for the hours where the data was missing; therefore, including zero would reduce the averages. Figure 2 – comparison of ANPR (Gantry 19) and CTO (Station 1894) data – and Figure 3 are two typical examples of comparative hourly flows for the ANPR and CTO data. Figure 2 data is typical of most of the ANPR/CTO comparisons, where flow profiles reveal only minor differences, with the ANPR data reflecting marginally higher average volumes. Figure 3, however, shows significant differences, where the ANPR data shows noticeably higher volumes. Investigating the differences revealed that lower CTO hourly averages resulted from missing data (zero) in the CTO database for periods of time. The cause of the missing data is unknown, but could be due to system malfunction. Since the ANPR data has no such data gaps, it can be considered at least as comprehensive and reliable a source of traffic volume data as the CTO systems, and in many cases better. As the ORT system is used to allocate toll transactions to road users, a high degree of accuracy and reliability is essential.

Added usefulness of ANPR data also depends on the ability to track vehicles
between camera locations through the recording and matching of number plates (VLN ID). The ANPR data from the gantries was processed to identify anomalies in terms of misread or otherwise unusable number plates. The information in Table 2 was provided by SANRAL's service provider for the electronic toll collection (ETC). These ANPR records were identified as either:

- Vehicles without number plates
- Unreadable number plates being obscured or damaged
- Illogical gantry combinations, possibly from cloned number plates passing gantries in illogical order or in impossibly short time periods.

These records cannot be used for number plate tracking, and effectively reduce the sample of the ANPR data for vehicle tracking by approximately 5%. This error was considered small enough that no correction or expansion of the remaining data was needed prior to its use for model validation. It is clear that, with a number of cases requiring a selection of the ANPR data for vehicle tracking.

### PROCESSING ANPR DATA AND OUTPUTS

Regarding the processing of the ANPR data, it must be noted that the ANPR cameras are on the toll gantries, and in this study reference to a gantry also means an ANPR camera location. Processing ANPR data for a selected time period required the development of a software program, which took the following into consideration:

- The traffic counts were to include all vehicles passing a selected gantry.
- All VLN anomalies (Table 2) were excluded from gantry-to-gantry (G2G) matrices.
- A maximum time needed to be specified to pass between adjacent gantries, before it is assumed that the vehicle left the freeway and re-joined it later to perform a second trip.
- A G2G distance matrix was derived from the gantry locations on the network.

Output from the software comprised traffic data relating to selected days of the week, times of the day and vehicle class.

The following traffic data was derived from the ANPR data:

- Hourly traffic flow profiles at each gantry, i.e. accurate traffic counts for the average week day and average day including weekends
- Average travel times between gantries for each hour of the day, which can be used to validate the modelled link volume delay curves on the freeway network
- Average speeds between gantries, which were calculated using the above G2G travel times and the G2G distance matrix
- Average G2G traffic counts, which are the numbers of vehicles that entered the freeway and were recorded passing a specified series of gantries within a specified time before leaving the freeway
- The trip length frequency distribution obtained by relating the G2G traffic counts to the G2G distance matrix. The G2G traffic counts are provided in matrix format, and Table 3 displays an example of the number of light vehicle trips between the first ten gantries (numbered in the first row and column) for the average weekday morning peak hour. Note that the full 42-gantry matrix has been reduced for clarity and the gantry numbers correspond to the gantry locations depicted in Figure 1.

The traffic counts on the diagonals represent the number of vehicles that enter and exit the freeway and only pass under the one gantry. In this matrix, these amount to about 64% of all observed trips, indicating a high usage of Gauteng freeways for short distance trips. The 219 trips from gantry 4 to gantry 6 enter the freeway between gantries 2 and 4 and exit the freeway between gantries 6 and 8. The downward trend in trips from gantry 2 to gantry 10, which are southbound trips on the N1, indicates a decreasing proportion of trips as the trip length increases. These G2G counts provide a independent data source to validate the distribution of trips that use the freeways in the traffic model. However, a methodology is required to extract comparative information from the traffic model.

### Table 2 ANPR records not used for vehicle tracking

<table>
<thead>
<tr>
<th>Record description</th>
<th>Number of records</th>
<th>Percentage of sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>No number-plate</td>
<td>1 397 571</td>
<td>1.8%</td>
</tr>
<tr>
<td>Unreadable/damaged number</td>
<td>1 127 586</td>
<td>1.4%</td>
</tr>
<tr>
<td>Illogical movements</td>
<td>1 016 415</td>
<td>1.3%</td>
</tr>
<tr>
<td>Total records not used for trips</td>
<td>3 541 572</td>
<td>4.5%</td>
</tr>
<tr>
<td>Total gantry passes</td>
<td>79 407 436</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

### Table 3 G2G light vehicle counts – weekday 07:00 to 08:00

<table>
<thead>
<tr>
<th>Light vehicles per hour</th>
<th>To gantry number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1 629</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>883</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>237</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>236</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>
**METHODOLOGY TO EXTRACT EQUIVALENT G2G TRAFFIC VOLUMES FROM A STRATEGIC TRAFFIC MODEL**

Select link (SL) analysis is a standard process incorporated in traffic modelling software that identifies the origins and destinations of all trips that use a certain link (Van Vliet 2015). Following the work of Carpenter *et al* (2012), select link (SL) analysis was used to derive an OD matrix for the trips that pass under each gantry.

Figure 4 depicts the possible trips that can be recorded through a notional freeway section with gantries (ANPR sites) A, B and C. The entry/exit points are numbered 1 to 14; these could be freeway, on-ramp or off-ramp nodes and represent the traffic model zones.

Let \( a \) denote the number of vehicle trips counted at gantry A. In model matrix format, the cells that contain trips through gantry A would include trips with origin zones 1, 2 and 3 and destination zones 4 to 14. Therefore, this includes trips that pass under gantry A only, under gantries A and B, and gantries A, B and C, and result in Select Link A (\( SL_A \)) matrix as shown in Matrix 1.

Similarly, the cells that contain trips included in a select link matrix through gantry B, will include vehicle trips (\( b \)) with origin zones 1 to 7 and destination zones 8 to 14, resulting in Select Link B (\( SL_B \)). Cells that contain trips that are included in a select link matrix through gantry C comprise trips (\( c \)) entering through zones 1 to 11 and exiting through zones 12 to 14. These trips include trips that pass under gantries A&B&C, B&C, and C only, resulting in Select Link C (\( SL_C \)). A combined select link analysis through gantry A, B or C results in \( SL_{ABC} \) as shown in Matrix 2.

Examining the matrices for \( SL_A \), \( SL_B \), \( SL_C \) and \( SL_{ABC} \) for a single cell, a trip that passes through gantries A, B and C in the three individual matrices, and the combined matrix is the same. Therefore:

\[
\begin{align*}
\text{Where cells contain: } & a, b & a = b \\
\text{Where cells contain: } & b, c & b = c \\
\text{Where cells contain: } & a, b, c & a = b = c \quad (1)
\end{align*}
\]

As a first step, to isolate the cells of an OD matrix which only relate to trips that pass through one “start” and one “end” location, A and B, and ignoring other gantries at this time, consider the combination of two

---

**Figure 4** G2G movements through three gantries

<table>
<thead>
<tr>
<th>Matrix 1</th>
<th>Select Link A – vehicle volumes (( a )) passing under gantry A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>2</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>3</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>4</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>5</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>6</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>7</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>8</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>9</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>10</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>11</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>12</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>13</td>
<td>a a a a a a a a a a a a</td>
</tr>
<tr>
<td>14</td>
<td>a a a a a a a a a a a a</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matrix 2</th>
<th>Combined trip matrix of vehicles passing gantries A, B or C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>a a a a a a b b a a b b a a a a</td>
</tr>
<tr>
<td>2</td>
<td>a a a a a a b b a a b b a a a a</td>
</tr>
<tr>
<td>3</td>
<td>a a a a a a b b a a b b a a a a</td>
</tr>
<tr>
<td>4</td>
<td>b b b b b b b b b b b b b b</td>
</tr>
<tr>
<td>5</td>
<td>b b b b b b b b b b b b b b</td>
</tr>
<tr>
<td>6</td>
<td>b b b b b b b b b b b b b b</td>
</tr>
<tr>
<td>7</td>
<td>b b b b b b b b b b b b b b</td>
</tr>
<tr>
<td>8</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>9</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>10</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>11</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>12</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>13</td>
<td>c c c c c c c c c c c c</td>
</tr>
<tr>
<td>14</td>
<td>c c c c c c c c c c c c</td>
</tr>
</tbody>
</table>
of the select link matrices $SL_A$ and $SL_B$ (Matrix 3). The cells of interest are only those that contain the trips $a,b$.

The next operation uses the Hadamard product (Horn & Johnson 2012), which simply multiplies the corresponding cells $(i,j)$ of two matrices of equal dimensions, i.e. $(X \cdot Y)_{ij} = (X)_{ij} \cdot (Y)_{ij}$. The Hadamard product of $SL_A$ and $SL_B$ will produce a zero where there is only $a$ or $b$, and $ab = a^2 = b^2$ in the cells containing $a,b$. The square root of the resulting cell elements will produce the matrix with all the cells that contain $a,b$ as shown in Matrix 4.

The sum of the model matrix trips that pass under gantry locations $A$ and $B$ can therefore be expressed as:

$$T_{G_{A,B}} = \sum_{i,j} [\sqrt{SL_A \cdot SL_B}]$$

(2)

Where:

- $T_{G_{A,B}}$ = the trips through gantry location $A$ and $B$
- $SL_A$ = Select Link matrix through gantry location $A$
- $SL_B$ = Select Link matrix through gantry location $B$

However, including gantry $C$, some trips that pass under gantries $A$ and $B$ also pass under gantry $C$, and Matrix 4 would include a $c$ in the cells representing origin zones 1, 2 and 3 and destination zones 12, 13 and 14. These trips should not be included in the desired result if only the trips between $A$ and $B$ and not through $C$, are required.

Subtracting $SL_C$ from Matrix 4 results in Matrix 5, since $c = a = b$. The desired trip matrix containing only those trips that pass under gantries $A$ and $B$ and not under gantry $C$ is obtained by removing the negative cells from Matrix 5, resulting in Matrix 6.

Summing the values in the resultant cells, which are the OD pairs of the trips that only pass under gantries $A$ and $B$, produces the equivalent of the G2G count in the G2G matrix from $A$ to $B$. This is given by Equation 3:

$$T_{G_{A,B}} = \sum_{i,j} [\sqrt{SL_A \cdot SL_B} - (\sum SL_C)]$$

(3)

Similarly, if one were to isolate the cells containing the trips that only pass through gantry location $B$, one can show that both $SL_A$ and $SL_C$ should be subtracted from the product. Equation 2 does not change, since both input matrices are $SL_B$. However, the resulting trip matrix contains trips through $A (a,b)$, $C (b,c)$ and $A$ and $C (a,b,c)$ as shown in Matrix 7.
Therefore, by summing all SL matrices identifying all possible alternative routes.

The process can, in general, be represented by the formula:

\[ T_{G_{AB}}^{\text{I/O}} = \sum_{T_{i,o}} \left( \sqrt{SL_{A} - SL_{B}} - (\sum SL_{A-1} + \sum SL_{B+1}) \right) \]  

(4)

Where:

- \( T_{G_{AB}}^{\text{I/O}} \) = trips from gantry A to gantry B only
- \( SL_{A} \) = Select Link matrix through gantry A
- \( SL_{B} \) = Select Link matrix through gantry B
- \( \sum SL_{A-1} \) = Select Link matrix(ies) of gantry(ies) upstream of gantry A
- \( \sum SL_{B+1} \) = Select Link matrix(ies) of gantry(ies) downstream of gantry B

The upstream and downstream gantries provide a "plug" on the ends of the desired section of the route. If there were more than one external gantry along the freeway, this should be added to the second part of Equation 4 above, i.e. replacing \( SL_{B+1} \) with \( (SL_{B+1} + SL_{B+2}) \). It was also discovered, while testing the formula on the GFIP model, that any route that provided a bypass to the first or last gantry A or B would "leak" traffic into the system from beyond the first or last gantry. A gantry that can be used as another external "plug" can be added to the external gantry list. This would, however, make the process onerous in a detailed network where there are potentially multiple alternative routes.

The solution to this problem lies in the fact that only positive cell values of the model’s SL matrices are added to derive the G2G equivalent value; therefore any number of other "plugs" can be added to the second part of Equation 3. It would also be easier to identify the ANPR camera (gantry) locations along a given route than identifying all possible alternative routes. Therefore, by summing all SL matrices \( SL_{ALL} \) and subtracting the sum of the SL matrices along the desired route, including \( SL_{A} \) and \( SL_{B} \) (i.e. \( SL_{B1} \)), the result would "plug" every possible "leak". Therefore, Equation 4 becomes:

Matrix 6 Model trip matrix containing trips that only travel between gantries A and B

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>2</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>3</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>8</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>9</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>10</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>11</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>12</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>13</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>14</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

Matrix 7 All trips included in \( SL_{B} \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>2</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>3</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
<td>a,b</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>8</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>9</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>10</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>11</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>12</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>13</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>14</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

Matrix 8 Subtraction of \( SL_{A} \) and \( SL_{C} \) from \( SL_{B} \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
</tr>
<tr>
<td>2</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
</tr>
<tr>
<td>3</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-a</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
<td>-c</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
</tbody>
</table>
$T_{G_{AB}} = \sum_{T_{ij}=0} (SL_A \cdot SL_B - (SL_{ALL} + SL_{RT}))$  \hfill (5)

Where:

- $T_{G_{AB}}$ = trips from gantry $A$ to gantry $B$ only
- $SL_A$ = Select Link matrix from the gantry $A$ link
- $SL_B$ = Select Link matrix from the gantry $B$ link
- $SL_{ALL}$ = sum of Select Link matrices of all gantry locations
- $SL_{RT}$ = sum of Select Link matrices along the route inclusive of gantry $A$ and gantry $B$

If trips between two gantries can choose two alternative routes, the ODs per route, or both routes, can be derived using:

- $SL_{RT_1}$ = sum of SL matrices along route 1
- $SL_{RT_2}$ = sum of SL matrices along route 2
- $SL_{RT_{1,2}}$ = sum of SL matrices along routes 1 and 2

**VALIDATING A TRAFFIC MODEL USING ANPR DATA**

Various data sets can be extracted from the ANPR data for use in validating a base year traffic model. As the ANPR data is not available for the base year itself (2006), the 2015 ANPR data from the ORT system was compared with the outputs of the 2015 forecasts from the GFIP traffic model. This serves to demonstrate the validation techniques described above.

**Traffic counts**

Figure 5 compares the modelled peak-period freeway traffic volumes at the gantry locations and the volumes derived from the ANPR data. These results show that, while the modelled and ANPR flows match quite well, the modelled light vehicle forecast is +9% too low, while the heavy vehicle forecasts are ±20% too low. A standard measure for comparing modelled volumes ($V_1$) and actual traffic volumes ($V_2$) in traffic modelling is the use of the GEH statistic represented by the following formula (Department for Transport 2014):

$$GEH = \sqrt{(V_2 - V_1)^2/(0.5(V_1 + V_2))}$$ \hfill (6)

The average GEH statistic across gantry locations was 8.04 for light vehicles. Whilst
not ideal (a validated model requires a GEH of 5 or less for 85% of observations), it must be noted that this is the comparison of a nine-year old forecast to measured counts and not the validation of a calibrated base year model; thus, some inaccuracy is to be expected. Under-forecasting in the model is most likely related to the high levels of toll non-payment experienced on the ORT network, as full toll compliance (assumed in the model) would have caused more deviation of trips from the tolled network to alternative roads.

**Journey times**

Journey times were extracted from the ANPR data, and checked for accuracy and consistency before being averaged for comparison to modelled journey times extracted from the traffic model. An example of the comparison for a section of the freeway spanning nine gantries is depicted in Figure 6. In this comparison, the modelled freeway journey time remains within the maximum recommended deviation of 15% from the measured journey times (Department for Transport 2014) over most of the length of the freeway. It also highlights specific freeway sections where the volume delay functions may require adjustment.

**Trip Length Frequency Distribution (TLFD)**

Figure 7 shows the comparison between the TLFD of the modelled trips using the freeways and the TLFD derived from the ANPR data using the G2G counts and distances between the gantries. Only light vehicles are shown for illustration. This correlation appears to validate the model in terms of the TLFD of trips using the freeways. It serves as an indication that the structure of the origin-destination matrix of freeway trips is close to accurate. The modelled average trip length for light vehicles, 11.43 km, is very close to the ANPR average of 11.30 km. However, as freeway trips are only a portion of all trips on the network, the same cannot be concluded for the model as a whole – additional trip length data from the remainder of the network is required to validate the rest of the network.

**Matrix trip distribution**

The G2G ANPR count matrix disaggregates 42 gantry counts into over 350 independent point-to-point counts with associated distances. Deriving an equivalent G2G matrix from the traffic model, using Equation 5 and the methodology described above, enables the validation of the distribution of trips within the cells of the model’s trip matrices. This comparison is shown in Figure 8. Whilst the TLFD from the ANPR data and the model correlate well, there is a greater variance in the disaggregated G2G counts, which in turn relates to the distribution of trips in the model trip matrices. On average, as can be expected, modelled values are slightly lower than actual ANPR values, the difference being about 6%.

Apart from providing an overall sense of the accuracy of the model’s OD matrix, the variances between individual cells can be examined to identify specific trends or problems. It was decided to focus only on the worst-performing gantry pairs as
No generally acceptable criterion exists for assessing OD volumes from a partial matrix, so the GEH statistic (Equation 6) was once again used to examine differences between the ANPR data and modelled values. The twelve gantry pairs with the highest (worst) GEH values are shown in Table 4.

Gantry numbers 8 and 32 are critical locations (refer to Figure 1) in that they are the entry arms to two of the highest trafficked system-interchanges on the network. With the “from” and “to” gantries being the same, it implies that these counts refer to short distance trips. These results therefore show that:

- Westbound on the N12 entering the Gillooly’s Interchange the model has nearly twice the number of short distance trips as ANPR.
- Southbound on the N1 entering the Buccleuch Interchange the model has approximately half the number of short distance trips as ANPR.
- Between gantries 19 and 21, i.e. southbound on the N3 travelling between the Buccleuch and Gillooly’s Interchanges, without using the N1 or N12, the model estimates over three times the number of actual trips.

This detailed comparison between the G2G counts and the model outputs highlights some significant localised discrepancies in the trip matrix distribution. This information can be very useful to pinpoint specific model improvements that may be needed, for instance where incorrect volume-delay curves were used in the freeway or (more likely) alternative route networks, leading to an incorrect assignment of trips onto the freeway. It is noted that this discrepancy could not have been picked up by only validating the model on the basis of the trip length frequency distribution, as the over- and under-assignments cancel each other out and leave the modelled TLFD close to the actual. It is the availability of large-scale ANPR data, and the partial OD matrix that results, which provide novel opportunities for matrix validation at levels of accuracy that were not possible before.

If a model is used to assess a scheme where a revenue stream or economic benefits are derived from distance-based costs and fares, the impact of these discrepancies may not be too significant if, as in the above case, the errors are averaged out in the TLFD. However, if no such averaging occurs, or if the model is to be used for a public transport scheme where the revenue is based on a boarding fare plus a distance-based fare, the number of short- and long-distance trips along the specific route can have a significant impact on the revenue stream. This revenue risk may have significant implications if the proposed scheme is part of a privately funded Public Private Partnership (PPP) concession (Bain 2009).

**SUMMARY**

When considering new large-data sources, one must identify the data’s strengths and weaknesses. ANPR data also has strengths and weaknesses in terms of all the data requirements of traffic models. Table 5 provides a summary of the traffic model data needs and ANPR’s strengths and weaknesses when compared to other large (electronically derived) data sources.

It is evident from the above that ANPR data is, like all other data sets, not the answer to all traffic model data needs. The major strength in the ANPR data is the ability to disaggregate the counts to independent counts over specific distances and enable the validation of a model’s trip distribution in the trip matrix. This has been enabled by the development of the methodology to extract equivalent count over distance (select link to select link) matrices from the model. The process of validating a traffic model using ANPR data has highlighted the fact that current

---

**Table 4 Twelve worst performing gantry pairs with highest GEH results based on ANPR and modelled G2G counts**

<table>
<thead>
<tr>
<th>Gantry From</th>
<th>Gantry To</th>
<th>ANPR</th>
<th>Model</th>
<th>% Diff</th>
<th>GEH</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>32</td>
<td>2513</td>
<td>4719</td>
<td>188%</td>
<td>36.69</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
<td>462</td>
<td>1583</td>
<td>343%</td>
<td>35.06</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>3084</td>
<td>1742</td>
<td>56%</td>
<td>27.32</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>397</td>
<td>1157</td>
<td>29%</td>
<td>27.25</td>
</tr>
<tr>
<td>31</td>
<td>31</td>
<td>1307</td>
<td>491</td>
<td>38%</td>
<td>27.21</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>341</td>
<td>0</td>
<td>0%</td>
<td>26.12</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>772</td>
<td>1598</td>
<td>207%</td>
<td>23.98</td>
</tr>
<tr>
<td>34</td>
<td>18</td>
<td>267</td>
<td>0</td>
<td>0%</td>
<td>23.12</td>
</tr>
<tr>
<td>34</td>
<td>20</td>
<td>264</td>
<td>0</td>
<td>0%</td>
<td>22.96</td>
</tr>
<tr>
<td>17</td>
<td>29</td>
<td>208</td>
<td>692</td>
<td>333%</td>
<td>22.8</td>
</tr>
<tr>
<td>32</td>
<td>20</td>
<td>397</td>
<td>998</td>
<td>25%</td>
<td>22.76</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>1208</td>
<td>555</td>
<td>46%</td>
<td>22</td>
</tr>
</tbody>
</table>

---

**Figure 8 Model validation by comparing individual OD counts between sets of gantries from ANPR data and modelled matrices**
methods of validating a traffic model may not uncover potentially critical problems in the distribution of trips in the matrix, even though the comparison to traffic counts and the TLFD show the model to be acceptable.

**FURTHER RESEARCH**

The extraction of the equivalent G2G counts from the model means that it is possible to produce a trip sub-matrix that only contains the trips that make up the G2G count. It is then possible to factor the cell values of the extracted sub-matrix so that the sub-matrix total equals the G2G count and re-inserting the sub-matrix values back into the original matrix. An iterative process of extraction, factoring and re-insertion would potentially improve the validation of the traffic model’s trip matrices, thus utilising the trip distribution characteristics of the ANPR data. This process is similar to current matrix estimation to traffic count techniques, except that there are more \( \frac{N(N-1)}{2} \) counts from \( N \) ANPR sites that are all independent – a desirable combination for matrix estimation (Ortúzar & Willumsen 1998).

**CONCLUSIONS**

The availability of ANPR data from the 201 km of the GFIP freeways utilising the 42 Open Road Tolling (ORT) gantries resulted in a significantly large data set, and an opportunity to assess this data for use in validating and improving traffic models. The processed data provided traffic counts, journey times along the freeway network and G2G (ANPR camera to camera) counts with related distances travelled on the freeways. Whilst the traffic counts and journey times provide dependable data, this information can be provided from other available means of data collection such as Comprehensive Traffic Observation (CTO) counting stations and journey time surveys from samples of probe vehicles.

The strength of the ANPR data lies in its ability to track large numbers of individual vehicles from point to point, thus producing counts over specific distances (G2G counts), which have the distribution of trips embedded in the data. The difficulty is that the G2G counts do not relate directly with the actual ODs in a model. The methodology developed to extract “G2G counts” from the traffic model has enabled the comparison of the ANPR data to the model outputs. From this work, the following can be concluded:

- ANPR technology can provide large accurate data sets that can be used for the development and validation of strategic traffic models. Where, as in the GFIP case, the ANPR data is intended for use in toll transactions, the coverage is near-complete in terms of vehicle volumes. Provided all lanes on the links are covered by the camera, there is no need to estimate expansion factors to represent the population.

- The location of the ANPR cameras can be either in a closed cordon or in an open layout, as in the GFIP freeway network. Optimising the location of the cameras would provide data sets with significant usefulness for both traffic modelling and traffic operation optimisation.

- As the ANPR data used in this research was limited to the freeway network, the traffic count and journey time data was also limited to the freeways, and hence a limitation in the ANPR data is that it only relates to a limited number of routes. Probe data (e.g. from on-board GPS equipment) has a broader coverage, and is useful for journey time information, even if smaller samples with unknown sample sizes are used.

- The major advantage of ANPR data is the ability to disaggregate the single point counts into accurate and independent counts associated with specific route distances, i.e. they can relate to the traffic model’s trip distribution.

- The method developed to isolate the trips in a traffic model’s trip matrix that represents the G2G counts enables the direct comparison of the ANPR data with the modelled trip distribution, hence offers a means to validate the partial OD matrix.

- The application of the above method to validate the GFIP model’s partial matrix showed that, even though a model’s journey times, TLFD and counts might be sufficiently accurate, there may still be irregularities in the trip matrices. The averaging of results may contribute to an acceptable validation outcome using standard validation procedures.

The comparison of partial OD matrices, based on ANPR data, may help to identify localised discrepancies in the trip matrix that can be very useful to pinpoint specific model improvements that may be needed, and that might otherwise be missed.

**ACKNOWLEDGEMENTS**

The authors would like to acknowledge the South African National Roads Agency SOC Limited for access to the ANPR data and the use of the Gauteng Freeway Improvement Project traffic model, as
well as Nicholas Robinson for the development of the software used to process the ANPR data.

REFERENCES


Estimating elastic moduli of sandstones using two-dimensional pore space images

TC Ekneligoda

The elastic moduli (shear modulus and bulk modulus) of two different sandstones are estimated from two-dimensional images of the pore space. An image analysis technique was used to extract the area and perimeter of each pore. The shearability and compressibility of each pore were calculated using the boundary element method and a perimeter-area scaling law. The effective shear modulus and bulk modulus of the rock were then estimated using the area-weighted mean pore shearability, pore compressibility and the differential effective medium theory. The method was applied to Fontainebleau and Berea sandstones. Comparison with experimental values of the shear moduli and bulk moduli showed good agreement.

INTRODUCTION
The macroscopic elastic moduli of a porous medium depend essentially on the porosity and the structure of the pores, along with a nearly trivial multiplicative dependence on the moduli of the non-porous host material. Methods for estimating the elastic moduli directly from images of the material would clearly be of great value compared to time-consuming and expensive experimental measurements. Numerous methods have been proposed to relate these microscopic elastic parameters that would occur in the “isolated pore or small porosity” limit to finite porosity limits (Christensen 1990; Zimmerman 1991; Grimvall 1999). However, no single method has gained universal acceptance.

Image analysis is a powerful tool, which is now widely used in several fields, including in engineering and medicine (Chermant 2001; Chermant et al 2001; Mouret et al 2001). In the recent past this technique has been successfully used in modelling the mechanical and transport properties of porous and inhomogeneous materials. Yue et al (2003) used the image analysis technique, together with the finite element method, to model asphalt concrete under different loading conditions. Chen et al (2004) used image analysis to aid in the prediction of inhomogeneous rock failure. Coster and Chermant (2001) discussed in detail how the technique can be incorporated into the modelling of civil engineering materials. Lock et al (2002) used information about rock pores obtained from image analysis of electron micrographs to predict the permeability of sandstones. In this paper, actual pore space images are used to study the relationship between the effective elastic properties and pore geometry. Image analysis is conducted to extract the pore geometry information, such as the perimeter and the area, from the pore images. The shearability parameter and compressibility parameter of the pores are estimated either by performing boundary element calculations, or using scaling laws. The differential effective medium theory is used to (approximately) account for the elastic interactions between nearby pores. This methodology is applied to Fontainebleau and Berea sandstones. The shear moduli and bulk moduli prediction are then compared with published experimental data.

PORE SHEARABILITY PARAMETER
The excess strain, $\Delta e$, due to an inhomogeneity in a homogeneous medium subjected to a far-field stress $\sigma^\infty$, is expressed in terms of the fourth-order $H$ tensor, as follows:

$$\Delta e = H: \sigma^\infty$$  \hspace{1cm} (1)

Where: the colon denotes the tensor inner product (Sevostianov et al 2008; Sevostianov & Kachanov 2002; Prokopiev & Sevostianov 2007).
For example, $H_{1212}$ connects the excess shear strain $\Delta \epsilon_{12}$ to the applied remote shear stress $\sigma_{12}$. For ellipsoidal pores, or special cases thereof, the $H$ tensor can be expressed in terms of the Eshelby tensor, of which the components are known (Eshelby 1957; Wu 1966).

The $H_{1212}$ component can be written to combine the effective shear modulus ($G$), as in Equation 2 (Ekneligoda & Zimmerman 2008). A similar relationship can be found (Drach et al 2011) which was used to evaluate the contribution of irregularly shaped three-dimensional pores to the overall elastic properties of carbon/carbon composites. Drach et al (2014) used a similar relationship to predict the effective elastic moduli of materials with irregularly shaped pores, based on the pore projected areas.

$$
\frac{1}{G_{eff}} = \frac{1}{G_a} + 2H_{1212}
$$

(2)

Where: $G_a$ is the shear modulus of non-porous or host material. Using the notation $H_{1212} = \phi S_{pc}$ ($\phi$ being porosity) hereafter, we use the term shearablity to denote $S_{pc}$.

For two dimensions, pores of a wide variety of shapes can be treated using the complex variable methods developed by Muskhelishvili (1963). For example, Ekneligoda and Zimmerman (2008) used this method to study the $H_{1212}$ of several quasi-polygonal pores, and other pores with $N$-fold axes of rotational symmetry. Their results showed that $H_{1212}$ for a planar strain condition, when the pore has $N$-fold rotational symmetry of order $N = 3, 5, 6$, etc, can be expressed as Equation 3 (note that $N$ cannot be equal to 4).

$$
H_{1212} = \frac{(k + 1)\phi}{2G_a [1 - m_1^2 (n - 2)] [1 - m_1^2 n]} (1 - m_1 \cos 4\theta)
$$

(3)

Where: $m_1$ and $n$ are the terms in the mapping function that conformally map the unit circle into the desired shape.

Detailed discussion about the limiting value of $m_1$ and $n$ is presented by Ekneligoda and Zimmerman (2008) – $k$ is $3 - \nu$ for plain stress, $G$ is the shear modulus and $\nu$ is the Poisson’s ratio.

Ekneligoda and Zimmerman (2008) explicitly analysed the fourfold ($N = 4 = (n + 1)$) condition. Under this condition an additional energy term arises. Their results showed that $H_{1212}$ for such pores that have fourfold ($N = 4$) can be expressed as:

$$
H_{1212} = \frac{(k + 1)\phi}{2G_a [1 - m_1^2 (n - 2)] [1 - m_1^2 n]} (1 - m_1 \cos 4\theta)
$$

(4)

Where: the angle $\theta$ is the anti-clockwise angle by which the stress state $\tau_{xy} = \tau_{yx} = \tau$ is rotated.

However, the use of complex variable methods to analyse the shearability parameter of an irregular pore that might be observed in a rock image is not practical, due to the fact that tens of thousands of terms would typically be needed in the mapping function that conformally maps the unit circle into the desired pore shape. Hence, it is important to find other, simpler methods to estimate the pore shearability.

One approach is to solve the equations of elasticity in the region outside of the pore using a numerical method such as the boundary element method (BEM). The boundary element calculations showed that, for irregular pores having no axis of symmetry, $H_{1212}$ always varies in the range 4 (Eroshkin & Tsukrov 2005). Hence, the only angular-dependent terms that can appear in $H_{1212}$ are those containing $\cos 4\theta$ and $\sin 4\theta$, or, equivalently $\cos(\delta + 4\theta)$, where the constant $\delta$ represents the phase shift. Alternatively, by retracing the steps that are involved in complex variable formulation, one can show that in the general case of an arbitrary number of terms in the mapping function, the angle $\theta$ enters the integral that expresses the excess energy only in the form of $e^{i4\theta}$. Hence, a $\cos(\delta + 4\theta)$ term should be expected to appear, but vanishes when the pore has $N$-fold rotational symmetry of order $N = 3, 5, 6$, etc.
In principle, the BEM method can be used to treat this problem. It would be simpler if the shearability parameter could be determined by geometric attributes of a pore shape. Ekneligoda and Zimmerman (2008) proposed the following scaling law to determine the shearability parameter of irregular pores:

\[
S_{PC} = \frac{GH_{121/2}}{2(1 – \nu)\phi} = 1 + \frac{1}{2} \left( \frac{P^2}{2\pi A} \right)
\]  

(5)

Where: \( P \) is the perimeter of the pore, \( A \) is its area, \( G \) is the shear modulus of the non-porous rock material, \( \phi \) is the porosity and \( \nu \) is the Poisson’s ratio of the rock material.

### Pore Compressibility

An important and conceptually simple parameter for use in discussing the elastic properties of porous media is the pore compressibility parameter, which can be defined as follows (Zimmerman 1991):

\[
C_{PC} = -\frac{1}{V_p} \left( \frac{\partial V_p}{\partial P} \right)_P
\]

Where: \( V_p \) is the initial pore volume, \( P \) is the external hydrostatic confining pressure, and \( V_p \) is the pressure of the pore fluid, which is held constant when taking the derivative described in Equation 6. The subscripts “p” and “c” refer to the compressibility of the pore with respect to the confining pressure.

The pore compressibility is of great importance in petroleum engineering, where it is used, for example, in material balance calculations. It is also important because it is used, for example, in material balance calculations. But it is also important because it is used, for example, in material balance calculations. But it is also important because it is used, for example, in material balance calculations. But it is also important because it is used, for example, in material balance calculations. But it is also important because it is used, for example, in material balance calculations.

The pore compressibility can be calculated for a pore of a given shape by solving the elasticity problem of an isolated pore in an infinite elastic solid. In three dimensions, this can be done analytically only for pores of the ellipsoidal family, which includes spheres, spheres, cylinders, and thin “penny-shaped” cracks. The pore compressibility of these shapes is essentially contained in the famous solution of Eschelby (1957). All other three-dimensional shapes require numerical solution (Burnley & Schmidt 2006). However, in two dimensions, pores of a wide variety of shapes can be treated using the complex variable methods developed by Muskhelishvili (1963). For example, Ekneligoda and Zimmerman (2006) used this method to study the compressibility of several quasi-polygonal pores, and other pores with n-fold axes of rotational symmetry.

Nevertheless, use of complex variable methods to analyse the compressibility of an irregular pore, such as might be observed in a rock image, is not practical, due to the fact that tens of thousands of terms would typically be needed in the mapping function that conformally maps the unit circle into the desired pore shape (Sisavath et al 2001). Hence, we must find other methods to estimate the pore compressibility. One approach is to solve the equations of elasticity in the region outside of the pore using a numerical method, such as the boundary element method (BEM).

The boundary element method can in principle be used to compute the compressibility of a pore of any shape, it would be simpler if the pore compressibility could be calculated from some simple geometric attributes of the pore shape, without requiring elaborate analytical or numerical calculations. Such a capability would be useful in attempting to establish a simple technique to estimate elastic moduli from images of heterogeneous media.

One approach to do this is the following scaling law that was proposed by Zimmerman (1986) for two-dimensional pores in plane-strain conditions:

\[
C_{PC} = \frac{2(1 – \nu)}{G} \frac{P^2}{4\pi A}
\]

Where: \( P \) is the perimeter of the pore, \( A \) is its area, \( G \) is the shear modulus of the non-porous rock material, and \( \nu \) is the Poisson ratio of the rock material.

This equation was “derived” by starting with the fact that an isolated circular pore has a compressibility of \( C_{PC} = (1 – \nu)/2G \) and then assuming that \( C_{PC} \) scales with the dimensionless ratio of perimeter-squared over area. The factor of 4π is needed to make the equation exact in the case of a circle.

Tsukrov and Novak (2002) tested this equation on a single, irregularly-shaped pore, and found that it was in error by only 8%. Ekneligoda and Zimmerman (2006) tested it for a variety of polygons and quasi-polygons, and found that the error was typically less than 10%, and never more than 21%.

Ekneligoda and Zimmerman (2006) showed that \( C_{PC} \) is always proportional to \((1 – \nu)/G\), with an additional dimensionless multiplicative factor that depends only on the pore shape. This factor can therefore depend on pore “shape” only through dimensionless parameters, such as \( P^2/A \).

However, this dependence is not necessarily linear, as is hypothesised in Equation 8.

Noting that the two-dimensional, plane-strain compressibility is related to \( G \) and \( \nu \) by \( C_{PC} = (1 – 2\nu)/G \), the proposed scaling law can also be written as:

\[
C_{PC} = \frac{2(1 – \nu)}{G} \frac{P^2}{4\pi A}
\]

### Effective Shear Modulus and Bulk Modulus

Consider Equation 2, with the \( S_{PC} \) term interpreted as pertaining to an isolated pore. We use the terms \( S_{PC} \) and \( S_0 \) to represent \( G_{eff} \) and \( G_o \) respectively. If we consider the thought experiment of inserting a small differential amount of pores into the non-porous rock, then one can say:

\[
S_{PC}(\delta\phi) = S_0 + S_{PC}\delta\phi
\]

(10)

A similar relationship can be built for pore compressibility as presented in Equation 11.

\[
C_{PC}(\delta\phi) = C_0 + C_{PC}\delta\phi
\]

(11)

Noting that \( S_0 \) is the effective shearability when the porosity is zero, one can convert Equation 10 into the following differential equation:
\[
\frac{1}{S_{\text{eff}}} \frac{dS_{\text{eff}}}{d\phi} = \frac{S_{\text{PC}}}{S_{\text{eff}}} \tag{12}
\]

Equation 12 can be used to model the evolution of the shearability, as additional pores are added into the rock, with the initial condition that \( S_{\text{eff}} = S_0 \) when \( \phi = 0 \).

Finally, it can be shown that Equation 12 can be simplified to Equation 13 with the reasonable assumption that \( S_{\text{PC}} \) to the host material \( S_{\text{eff}} \) is constant with the addition of new pores. In this case, its value must be equal to the value it had in the isolated pore, zero-porosity limit. We call this value \( S_{\text{PC}}/S_0 \), with the superscript "o" used to denote the zero-porosity limit.

\[
\frac{S_{\text{eff}}}{S_0} = \exp \phi \left[ \frac{S_{\text{PC}}}{S_0} \right] \tag{13}
\]

Following the similar argument, the variation of compressibility parameter can be expressed as:

\[
\frac{C_{\text{eff}}}{C_o} = \exp \phi \left[ \frac{C_{\text{PC}}}{C_o} \right] \tag{14}
\]

In the above scheme it was implicitly assumed that, as new pores with incremental porosity \( \delta\phi \) are placed into the material, the total porosity is increased by \( \delta\phi \). But Norris (1985) and others have pointed out that, if these additional pores are imagined to be randomly placed in the rock, they will replace solid material with probability \( 1 - \phi \), and replace existing pores with probability \( \phi \). Hence, the increment in total porosity will be \( \delta\phi/(1 - \phi) \). If we replace \( \delta\phi \) in Equation 13 with \( \delta\phi/(1 - \phi) \), this has the effect of transforming Equation 13 into Equation 15:

\[
\frac{S_{\text{eff}}}{S_0} = (1 - \phi) \frac{S_{\text{PC}}}{S_0} \tag{15}
\]

In terms of the effective shear modulus, Equation 15 takes the form:

\[
\frac{G_{\text{eff}}}{G_o} = (1 - \phi) \frac{S_{\text{PC}}}{S_0} \tag{16}
\]

Following the similar approach the relationship for bulk modulus can be expressed as:

\[
\frac{K_{\text{eff}}}{K_o} = (1 - \phi) \frac{C_{\text{PC}}}{C_o} \tag{17}
\]

This is the same form, \((1 - \phi)^m\), that has often been proposed empirically to model the elastic moduli of porous ceramics (Rice 1998), but the present derivation assigns a physical interpretation to the exponent \( m \). All that remains is to estimate the value of the shearability and compressibility parameters, which depend on pore shape.

**IMAGE ANALYSIS AND NUMERICAL CALCULATION TO DETERMINE THE PORE SHEARABILITY AND PORE COMPRESSIBILITY**

We start with a scanning electron micrograph (SEM) image of a region of rock surface. Figure 2(a) shows such an image for a Berea sandstone, taken from Schlueter et al (1997). The dark grey regions are quartz grains. The pore space is impregnated with Wood’s metal (white) and epoxy (black), and the light grey regions are clays and other minerals. This helped to identify the total pore space. This image was initially taken for a study of two-phase flow properties, but for the present purposes the only important distinction to be made is between pores and minerals. The size of the original image in Figure 2(a) is 375 × 250 μm.

The image analysis software packages Idrisi and Carta-Linx are used to extract the pores from the images, and compute their associated properties, such as area and perimeter. As the first step the original image is enhanced, using 4 × 4 kernel with mode filtering, to arrive at Figure 2(b). Then, based on the histogram of colour.
distribution, a binary image was created to separate the pores from the minerals, as shown in Figure 2(c). As an intermediate step, smaller features of size less than 8 μm² were eliminated, using a histogram of pore areas. These features may simply be artifacts of the image analysis procedure, or may indeed be small pores. With regard to the latter possibility, we note that these smaller pores contribute only 2.5% to the total porosity of the analysed section.

Finally, a vector image is created in which the boundary of each pore is associated with a numerical (x,y) coordinate pair, as shown in Figure 2(d).

To remove artificially introduced waviness in the pore boundary, the vector image was exported to another image analysis software, Carta-Linx. Smoothing was carried out using the option called “generalisation”, based on the individual perimeter values of the pores. This is permissible, as waviness with small amplitudes does not contribute any additional stiffness to the pore. The entire process of image analysis can be summarised as presented in Figure 3.

Images of two Fontainebleau sandstone samples, having porosities of 12% and 25%, are shown in Figures 4 and 5. These images illustrate the fact that essentially any two-dimensional pore image can be used as a starting point for the procedure, as long as the pores can be distinguished from the minerals. The image analysis procedure described above was applied to the Fontainebleau sandstone samples shown in Figures 4 and 5.

The boundary element method is then used to calculate $S_{pc}$ and $C_{pc}$ numerically, assuming each pore to be isolated in an infinite rock matrix. The average value of $S_{pc}$ is determined after the full rotation of stress as represented in Figure 1 (i.e $\theta$ was varied from 0° to 360° at steps of 10°). The calculations are performed using the code developed by Martel and Muller (2000).

In the calculations of $C_{pc}$, all farfield stresses and body forces are set to zero. A uniform normal traction of unit magnitude is prescribed over the surface of the hole to determine the area change of the single pore. The cavity boundary is discretised into a number of equal-length elements. We generally found that roughly 300 boundary elements are sufficient to achieve convergence of the computed compressibilities.

RESULTS

Berea sandstone

The area and perimeter of each individual pore from the Berea sandstone image in

<table>
<thead>
<tr>
<th>Perimeter (μm)</th>
<th>Area (μm²)</th>
<th>$G_{pc}((1 - v)/BEM)$</th>
<th>$G_{pc}((1 - v) (Eq 3))$</th>
<th>Error of (Eq 3) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>293</td>
<td>4 369</td>
<td>2.58</td>
<td>3.12</td>
<td>21</td>
</tr>
<tr>
<td>475</td>
<td>9 080</td>
<td>3.39</td>
<td>3.96</td>
<td>17</td>
</tr>
<tr>
<td>475</td>
<td>9 080</td>
<td>3.02</td>
<td>3.71</td>
<td>8</td>
</tr>
<tr>
<td>251</td>
<td>12 510</td>
<td>3.99</td>
<td>4.90</td>
<td>23</td>
</tr>
<tr>
<td>331</td>
<td>3 025</td>
<td>5.93</td>
<td>5.76</td>
<td>3</td>
</tr>
<tr>
<td>259</td>
<td>3 734</td>
<td>2.63</td>
<td>2.87</td>
<td>9</td>
</tr>
<tr>
<td>728</td>
<td>12 476</td>
<td>6.03</td>
<td>6.76</td>
<td>12</td>
</tr>
<tr>
<td>240</td>
<td>2 803</td>
<td>3.33</td>
<td>3.27</td>
<td>2</td>
</tr>
<tr>
<td>173</td>
<td>3 637</td>
<td>2.44</td>
<td>2.54</td>
<td>4</td>
</tr>
<tr>
<td>210</td>
<td>1 879</td>
<td>2.61</td>
<td>2.65</td>
<td>2</td>
</tr>
<tr>
<td>326</td>
<td>2 657</td>
<td>3.29</td>
<td>3.46</td>
<td>5</td>
</tr>
<tr>
<td>309</td>
<td>4 889</td>
<td>4.56</td>
<td>5.15</td>
<td>13</td>
</tr>
<tr>
<td>157</td>
<td>1 282</td>
<td>3.10</td>
<td>3.04</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 2(a) are shown in Table 1, along with the pore shearability that is calculated by the boundary element method (BEM), and estimated from scaling law (Equation 3). The same information is also plotted in Figure 6.

Table 2 summarises the pore compressibility parameter derived based on scaling law and the boundary element method. Figure 7 illustrates the compressibility parameter derived by scaling law and boundary element methods. The maximum shearability parameter is 4.4 and the maximum compressibility parameter becomes 6.8 in the analysed Berea sandstone section.

The largest pore analysed in the section had an area of 12,476 μm² and a perimeter of 728 μm. Only two pores gave an error percentage larger than 10%.

The maximum error percentage was given by the largest pore in the analysed section which had an area of 12,509 μm² and a perimeter of 621 μm. However, the error percentage of the shearability of the same pore is 6.48. The pore that resulted in the second-highest error percentage of shearability, however, gave only 3% error in the compressibility parameter.

The porosity of the Berea sandstone in Figure 2(a) is 0.22, so Equation 16 gives an effective shear modulus of 14.5 GPa. The mean value of $G_oC_{pc}/(1 - \nu_o)$, as calculated using the scaling law, is 3.53, in which case the calculation procedure yields $G_{eff} = 13.0$ GPa. These compare very well with the high-stress effective shear modulus of 14.1 GPa measured by King (1966).

The area-weighted mean value of the normalised pore compressibility $G_oC_{pc}/(1 - \nu_o)$, as calculated by the BEM, is 3.96. The elastic moduli of the mineral phase of Berea, as reported in Zimmerman (1991) are $G_o = 31.34$ GPa, $K_o = 39.75$ GPa, and $\nu_o = 0.188$. Hence, we see that $C_{pc}/C_o = 4.58$. The porosity of Berea sandstone is 0.22, so Equation (17) gives an effective bulk modulus of 13.0 GPa. The mean value of $G_oC_{pc}/(1 - \nu_o)$, as calculated by the scaling law, is 4.44, in which case the analogous calculation procedure yields $K_{eff} = 11.1$ GPa. These values are both reasonably close to the experimental value, measured at high stresses when all microcracks are closed, which is 9.6 GPa (Zimmerman 1991).

Fontainebleau sandstone

The area-weighted pore shearability of the sample with $\phi = 12\%$ (Figure 4) is found to be $G_oS_{pc}/(1 - \nu_o) = 2.58$ according to the BEM calculations, and 3.50 according to
the scaling law. For the $\phi = 25\%$ sample (Figure 5) these values were 3.33 and 3.48 respectively. The comparison of pore shearability parameters of individual pores for porosity 25% sample determined using the scaling law and the numerical method is presented in Figure 8. Interestingly, the pore shearability values do not seem to vary very much with porosity.

The value computed for the 25% porosity sample is used in the following calculations: The elastic moduli of the mineral phase of Fontainebleau sandstone, which is essentially pure quartz, are $G_o = 44$ GPa, $K_o = 37$ GPa and $\nu_o = 0.074$ (Arns et al 2002). It is already stated that $S_{pc}/S_o = 3.33$ according to the BEM calculations, and $S_{pc}/S_o = 3.48$ using the scaling law. Finally, Equation (16) yields the following prediction for the effective shear modulus, when the pore shear compliance is calculated by BEM:

$$G_{eff}(\text{GPa}) = 44.0(1 - \phi)^{3.33}$$

These two effective shear modulus predictions are shown in Figure 9, where they are compared to the experimental values reported by Arns et al (2002), and to the three predictions made by Prokopiev and Sevostianov (2007). As was the case for the bulk modulus, the present method gives very good predictions over the entire range of data. Note that the values predicted by the finite element calculations (FEM) of Arns et al (2002) are indistinguishable from the predictions of Equation (16), in which the individual pore compliances were found from BEM calculations, and the differential effective medium method was used to extrapolate from the low-porosity limit out to higher porosities.

The same image analysis procedure as was described above for Berea is then applied to the Fontainebleau images. The area-weighted pore compressibility of the sample with $\phi = 12\%$ is found to be $G_o C_{pcc}/(1 - \nu_o) = 3.81$ according to the BEM calculations, and 5.02 according to the scaling law. For the $\phi = 25\%$ sample, these values were 3.84 and 4.97 respectively. Interestingly, the pore compressibility values do not seem to vary much with porosity; we will use the values computed for the 12% porosity sample in the following calculations. The elastic moduli of the mineral phase of Fontainebleau, which is essentially pure quartz, are (Arns et al 2002) $G_o = 44$ GPa, $K_o = 37$ GPa and $\nu_o = 0.074$. Hence, we find $C_{pc}/C_o = 2.97$ according to the BEM calculations, and

![Figure 8](image_url) Test of shear compliance scaling law (Equation 3) for the Fontainebleau sandstone pores shown in Figure 2; values computed with BEM are plotted on the x-axis, predictions of the scaling law are plotted on the y-axis.

![Figure 9](image_url) Shear modulus of Fontainebleau sandstone, showing values measured by Arns et al (2002), along with those predicted by the present methods, by the finite element calculations of Arns et al (2002) and Prokopiev and Sevostianov (2007).
and 21% for Berea sandstone. The variation of pore compressibility from 12% porosity to 22% porosity is less than 1% for the boundary element calculation and scaling law for the analysed Fontainebleau sandstone. The resulting predictions are close to the shear moduli and bulk moduli values measured in the laboratory.

Our methodology therefore offers a promising and simple approach to relate the mechanical properties of porous geological materials to their pore structure, using only two-dimensional images.

ACKNOWLEDGEMENT

The author thanks Prof Robert Zimmerman for his valuable support during the preparation of this paper.

REFERENCES


Guidelines for the preparation of papers and technical notes

Authors should comply with the following guidelines when preparing papers for publication in the journal.

The Journal of the South African Institution of Civil Engineering is published quarterly in March, June, September and December. Articles submitted for publication are reviewed by a panel of referees under the guidance of the SAICE Journal Editorial Panel. The journal publishes research papers covering all the disciplines of civil engineering (structural, geotechnical, railway, coastal/marine, water, construction, environmental, municipal, transportation) and associated topics that are relevant to the civil engineering profession, and that preferably have to do with civil engineering in southern Africa and the African continent.

When preparing articles for publication, authors should please take note of the following and comply with the guidelines as set out:

CLASSIFICATION OF ARTICLES CONSIDERED FOR PUBLICATION

- Technical papers are well-researched, in-depth, fully referenced technical articles not exceeding 6 000 words in length (excluding tables, illustrations and the list of references). Related papers that deal with ‘softer sciences’ (e.g. education, social upliftment, etc) are accepted if they are of a technical nature and of particular interest to the civil engineering profession. The latter type of paper will be subject not only to peer-review by civil engineers, but also to review by non-engineering specialists in the field covered by the paper.

- Technical notes are short, fully referenced technical articles that do not exceed 2 000 words. A typical technical note will have limited scope often dealing with a single technical issue of particular importance to civil engineering.

- Review papers are considered for publication as either technical papers or technical notes on condition that they are the original work of the author and will assist the reader with the understanding, interpreting or applying of the subject under review. A review paper must contain criteria by which the work under review was evaluated, and contribute by synthesising the information and drawing new conclusions from the dissemination of the previously published work.

- Discussion on published articles is welcomed up to six months after publication. The length of discussion contributions is limited to 1 500 words. Where appropriate, discussion contributions will be subject to the normal reviewing process and will be forwarded to the authors of the original article for reply.

POLICY REGARDING LANGUAGE AND ORIGINALITY OF SUBMITTED ARTICLES

- Language: Manuscripts should preferably be presented in English, as the journal is distributed internationally. Articles submitted in any of the other official South African languages should be accompanied by an expanded abstract in English.

- Original work: Papers and technical notes must be original contributions. Authors must affirm that submitted material has not been published previously, is not under consideration for publication elsewhere and will not be submitted elsewhere while under consideration by the SAICE Journal Editorial Panel. It is the responsibility of the authors to ensure that publication of any paper in the journal will not constitute a breach of any agreement or the transgression of any law. The corresponding author should confirm that all co-authors have read and approved the manuscript and accept these conditions. Authors are responsible for obtaining permission to publish experimental data and other information that may be confidential or sensitive. Authors are also responsible for obtaining permission from copyright owners when reproducing material that had been published elsewhere. Proof of such permission must be supplied.

SUBMISSION PROCEDURES AND REQUIRED FORMAT

- Online submission: Manuscripts must be uploaded as PDF files (http://journal.saice.org.za). Individual file sizes may not exceed 10 MB. Should you experience problems uploading your paper, please contact the editor (verelene@saice.org.za).

- Format: Manuscripts should be prepared in MS Word and presented in double line spacing, single column layout with 25 mm wide margins. Line numbers must be applied to the whole document. All pages should bear the authors’ names and be numbered at the bottom of the page. With the exception of tables and figures (see below) the document should be typed in Times New Roman 12 pt font. Contributions should be accompanied by an abstract of not more than 200 words.

- First page: The first page of the manuscript should include the title of the paper, the number of words of the main text (i.e. excluding figures, tables and the list of references), the initials and surnames of the authors, professional status (if applicable), SAICE affiliation (Member, Fellow, Visitor, etc), telephone numbers (landline and mobile), and e-mail and postal addresses. The name of the corresponding author should be underlined. Five keywords should be suggested.

- Figures, tables, photos and illustrations: These should preferably be submitted in colour, as the journal is a full-colour publication.

- Their positions should be clearly marked in the text as follows: [Insert Figure 1].

- Figures, tables, photos, illustrations and equations should be numbered consecutively and should appear in the text directly after they have been referred to for the first time.

- Illustrations must be accompanied by appropriate captions. Captions for tables should appear above the table. All other captions should appear below the illustration (figures, graphs, photos).

- Only those figures and photographs essential to the understanding of the text should be included. All illustrations should be referred to in the text.

- Figures should be produced using computer graphics. Hand-drafted figures will not be accepted. Lettering on figures should be equivalent to a Times New Roman 9 pt font or slightly larger (up to 12 pt) if desired. Lettering smaller than 9 pt is not acceptable.

- Tables should be typed in Times New Roman 9 pt font. They should not duplicate information already given in the text, nor contain material that would be better presented graphically. Tabular matter should be as simple as possible, with brief column headings and a minimum number of columns.

- Mathematical expressions and presentation of symbols:

- Equations should be presented in a clear form which can easily be read by non-mathematicians. Each equation should appear on a separate line and should be numbered consecutively.

- Symbols should preferably reflect those used in Microsoft Word Equation Editor or Mathtype, or should be typed using the Times New Roman symbol set.

- Variables in equations (e.g. x, y, z, etc) and lower case Greek letters) should be presented in italics. Numbers (digits), upper case Greek letters, symbols of metric measurement units (m for metres, s for seconds, etc) and mathemati-cal/trigonometrical functions (such as sin, cos and tan) are not written in italics, but in upright type (Roman). Variables and symbols used in the body of the text should match the format used in the equations, i.e. upright or italics, whichever is applicable.

- Metric measurement abbreviations/units should conform to international usage – the SI system of units should be used.

- Decimal commas may be used, but decimal points are preferred.

- Symbols should preferably be defined in the text, but if this is not feasible, a list of notations may be provided for inclusion at the end of the paper.

- Headings: Sections and paragraphs should not be numbered. The following hierarchy of headings should be followed:

  - HEADING OF MAIN SECTION
  - Heading of sub-section
  - Heading of sub-sub-section

- References: References should follow the Harvard system. The format of text citations should be as follows: “Jones (1999) discovered that…” or “recent results” (Brown & Carter 1985; Green et al 1999) indicated that “…” References cited in the text should be listed in alphabetical order at the end of the paper. References by the same author should be in chronological order. The following are examples of a journal article, a book and a conference paper:


- Papers published previously in the Journal of the South African Institution of Civil Engineering should be cited if applicable.

- Footnotes, trade names, acronyms, abbreviations: These should be avoided. If acronyms are used, they should be defined when they first appear in the text. Do not use full stops after abbreviations or acronyms.

- Return of amended papers: Papers requiring amendments will be accepted up to six months after the referee reports had been sent to authors, after which the paper will be withdrawn from the system.

FINAL ARTICLE

- Copyright: On acceptance of the paper or technical note, copyright must be transferred by the author(s) to the South African Institution of Civil Engineering on the form that will be provided by the Institution.

- Photos of authors: The final corrected version of the paper should be accompanied by recent, high-resolution head and shoulders colour photographs and a profile not exceeding 100 words for each of the authors.

- Proofs: First proofs of papers will be sent to authors in PDF format for verification before publication. No major re-writes will be allowed, only essential minor corrections.