Title: Application of machine learning techniques for asphaltic material modeling

Candidate: Abhary Eleyedath

ABSTRACT

It is well known fact that determination of properties of asphaltic materials in laboratory is laborious, expensive, time and resource consuming exercise. Hence, researchers have resorted to use of tools that provide a meaningful predictive model for material response. Within the domain of asphaltic materials, there are very few models available for predicting the material response. Owing to predictive capability, researchers/engineers have resorted to use of machine learning schemes to develop predictive models. However, most of the approaches/schemes are black box in nature. Under these circumstances, Gene Expression Programming (GEP) comes handy. GEP is a non-parametric technique that uses a symbolic regression approach to optimise the function itself. One of the main advantage of this technique is availability of equation that can be used as predictive model. The present work proposes GEP based models for predicting response of different asphaltic materials. These include (i) prediction of effective viscosity of binary blends of asphalt binders, (ii) prediction of density and viscosity of heavy crudes, (iii) prediction of foamed bitumen properties, and (iv) dynamic modulus of asphalt concrete. Several improvements have been suggested to improve prediction capability of GEP based approach. The same has been discussed in following paragraphs.

The first objective was to develop a viscosity mixing rule for binary blend of asphalt binders using GEP. This work presents a novel GEP based approach to obtain a viscosity-mixing rule. To develop these expressions, two distinct binary blends comprising of varying proportion of unmodified binder and polymer modified binder were prepared and tested for their resultant viscosity at different temperatures. The obtained data were used to (i) develop the GEP-based viscosity-mixing rule, and (ii) compare the resultant viscosity with viscosity-mixing rules reported in the literature (like the Arrhenius model). Statistical analysis indicated that the accuracy of GEP based mixing rule is superior over other viscosity-mixing rules reported in the literature.

The prediction of density and viscosity of heavy bitumen using GEP was chosen as the second objective. To evaluate the accuracy of proposed GEP based models, results reported by various researchers were utilized. Published literature on heavy bitumen including Athabasca, Cold lake and Gas free bitumen were used for this purpose. Using temperature, and pressure

data as input, the density and viscosity of bitumen were predicted using GEP technique. The developed GEP based models were compared with the conventional empirical regression equations proposed by others. The statistical analysis indicated that GEP based models work better than currently used models for density and viscosity of bitumen.

The foamed bitumen parameters (i.e., ER, HL) are highly influenced by viscosity, foaming temperature and water content. With no predictive model that relates these variables with foaming parameters, development of predictive model was taken as third objective. The database consisting of ER and HL, and physico-chemical properties of six distinct binders was developed by research team at Central Road Research Institute (CRRI), New Delhi. All binders were tested for their physical properties and chemical composition, and the ER and HL were measured over the range of test temperatures (110°C to 200°C) and water contents (2% to 11%). The final database consisted of 166 observations. Expressions to predict ER and HL have been developed based on temperature, water content, viscosity, and chemical composition using this database through the GEP technique. The goodness of fit parameters indicated that models based on physical properties are able to predict ER and HL with reasonable accuracy and with the addition of chemical composition along with physical properties improved the accuracy of predictive models significantly. These expressions can be used to identify the feasible range of test conditions for the production of foamed bitumen before actual testing is commenced.

The fourth objective was to develop a novel hybrid clustering- GEP approach for predicting foamed bitumen properties. A database consisting of 190 observations was used. Essentially this database was obtained by addition of more binder data to database used with third objective. The Self-Organizing Map (SOM) based clustering of this database helped in obtaining homogeneous groups even with highly complex interaction. Further, C5.0 algorithm was used to decipher underlying patterns among clusters identified by SOM. GEP approach was used to develop four global models to predict HL and ER. Subsequently hybrid models were obtained through recalibration of these global models but using data from individual clusters. These models were different than those obtained while working on third objective. Major differences include (i) different functional form, (ii) additional model for HL in terms of ER, and (iii) increased accuracy with more information supplied. Statistical analysis indicated that hybrid models outperformed corresponding global models in all cases. Global sensitivity analysis indicated that among various parameters, water content had significant effect on ER

prediction. This was followed by temperature, and viscosity**.** However, for predicting HL, this order was ER (if used), water content, temperature, and viscosity.

The fifth objective of research presents a novel hybrid Principal Component Analysis (PCA) - GEP approach to predict the $|E^*|$ of asphalt concrete. For this purpose, $|E^*|$ database developed during NCHRP 9-19 study was used. Using the information of all properties as input (i.e., variables), the dimensionality was reduced using PCA. Such an exercise helped in removing the redundancy at input stage. The extracted principal components were used to develop first set of $|E^*|$ prediction models. The feature selection property of PCA was used to decide the parameters mostly contributing to the individual principal components (PC's) and rank the same. Using the ranked variables as input, second set of $|E^*|$ prediction models were developed. Careful analysis of these two sets of models indicated that addition of parameters (or PC's) beyond certain number did not contribute to accuracy of model. Comparison of models from these two sets indicated that predictive model obtained using variables as direct input resulted in improved accuracy. For better understanding, this finalized model was compared with other regression-based equations proposed by others. Comparison of goodness of fit indicators indicated that proposed hybrid model offers efficient and accurate alternative. The proposed model has flexibility to be used with any new database with recalibration.

The sixth (and final) objective of work evaluates the underlying patterns in prediction error that exists in these models. Again $|E^*|$ database developed during NCHRP 1-40D study was used for this purpose. Initially, the global dataset was compared against individual mixtures for similarity in terms of mean, range, and correlation. T-test showed that the individual mixture datasets are significantly different from the global dataset. Thus, calibration of predictive models was undertaken using global dataset and individual mixture dataset. Statistical indicators indicated improvement with mixture-wise calibration when compared to global calibration. To check the extent of prediction error, 'difference parameter' (DP) and 'ratio parameter' (RP) was introduced in this work. Q-Q plots and cumulative distribution plots constructed using DP and RP indicated highest and lowest error with Al-Khateeb and PCA-GEP models, respectively. For a detailed analysis, the entire range of measured $|E^*|$ was divided into subdivisions. In general, lower prediction error was observed in middle range of measured $|E^*|$. All the predictive models considered in this study overpredicted and underpredicted in lower and higher modulus range, respectively. Also, in the lower and higher modulus region, DP and RP showed skewed and flatter distribution when compared to normal distribution. Based on the numerical values of these distribution indicators, it was inferred that performance of all models are comparable in middle $|E^*|$ range. However, in extreme values of $|E^*|$, Hirsch and Al-Khateeb models performed poorly. In the same range, PCA-GEP, Original Witczak and South Korean models performed well. Sensitivity analysis indicted that binder properties exhibited highest sensitivity followed by testing condition, aggregate gradation and volumetrics.