1 **Comparison of Different Data Mining Techniques for Predicting Compressive Strength of**

2 **Environmentally Friendly Concrete**

3 Behzad Abounia Omran[1]; Qian Chen, A.M.ASCE[2]; and Ruoyu Jin[3]

4 **Abstract:** With its growing emphasis on sustainability, the construction industry is increasingly

5 interested in environmentally friendly concrete produced using alternative and/or recycled waste

6 materials. However, the wide application of such concrete is hindered by lack of understanding

7 of the impacts of these materials on concrete properties. This research investigates and compares

8 the performance of nine data mining models in predicting the compressive strength of a new type

9 of concrete containing three alternative materials as fly ash, Haydite® lightweight aggregate, and

10 Portland limestone cement. These models include three advanced predictive models (multilayer

11 perceptron, support vector machines, and Gaussian processes regression), four regression tree

12 models (M5P, REPTree, M5-Rules, and decision stump), and two ensemble methods (additive

13 regression and bagging) with each of the seven individual models used as base classifier. The

14 analytical results show that with appropriate parameter settings all of these models except for

15 decision stump achieved acceptable prediction performance. The ensemble methods improved

16 the prediction accuracy of the four regression tree models, but had less success on the other three

17 advance predictive models. The individual Gaussian processes regression model as well as its

18 related ensemble models reached the highest prediction accuracy in comparison groups. The

19 results of this paper offer valuable insights to improving the use of these models for property

20 prediction of concrete.

[1] Ph.D. Candidate, Construction Systems Management, Dept. of Food, Agricultural, and Biological Engineering, The Ohio State University, 590 Woody Hayes Dr., Columbus, OH 43210. E-mail: abounia-omran.1@osu.edu.

[2] Associate Professor, Construction Systems Management, Dept. of Food, Agricultural, and Biological Engineering, The Ohio State University, 590 Woody Hayes Dr., Columbus, OH 43210 (corresponding author). E-mail: chen.1399@osu.edu.

[3] Assistant Professor, Dept. of Architecture and Built Environment, University of Nottingham Ningbo China, 323 Science and Engineering Building, 199 Taikang East Rd., Ningbo, 315100, China. This work was conducted while he was a graduate research assistant at The Ohio State University. E-mail: Ruoyu.Jin@nottingham.edu.cn.

1

21   **CE Database subject headings:** Construction material; Concrete; Compressive Strength, Data

22   Analysis, Predictions, Sustainable Development; Neural network; Gaussian processes.

23   **Author keywords:** Machine learning; Data mining; Predictive models; Environmentally friendly

24   concrete; Comparison.

Introduction

The construction industry has observed an increasing shift toward sustainability in recent years. Many companies are proactively using or are required by their clients to use more environmentally friendly (or so-called "green") building materials and/or processes to reduce the environmental effects from construction activities. Environmentally friendly concrete is defined as concrete produced using alternative and/or recycled waste materials that can lower the overall environmental impacts of concrete during its life cycle. This type of concrete increasingly becomes a common element that helps the construction industry achieve long-term sustainability, although the impact of these alternative or recycled waste materials on various concrete properties has not been fully understood.

Using alternative materials in concrete may positively or negatively impact its properties (Khalaf and Devenny 2004; Yang et al. 2005; Berry et al. 2011). Research is thus needed to thoroughly understand the potential influence from these materials. Since the compressive strength is one of the most important concrete properties, many experiments have been conducted to study the compressive strength of environmentally friendly concrete (Yang et al. 2005; Etxeberria et al. 2007; Kevern et al. 2011). Despite some progress, the available data for such concrete is far from adequate due to the emergence of so many alternative or recycled waste materials and the complexity of concrete mixture design. Not only is more research needed to advance the understanding of environmentally friendly concrete properties, but practical tools for designing these types of concrete are necessary for wide implementation.

Differing from the traditional experimental method, some researchers proposed mathematical or statistical models to predict the compressive strength of concrete given its mixture or based on fresh concrete properties (Atici 2011). Statistical modeling has its

48 limitations in estimating the underlying relationships between the inputs and outputs of

49 forecasting models in more complicated cases (Zhang 1998). As a result, recent studies have

50 shown an increasing trend toward the application of machine learning techniques in predicting

51 concrete compressive strength (Topçu and Saridemir, 2007; Saridemir et al. 2009; Atici 2011;

52 Aiyer et al. 2014; Akande et al. 2014; Omran et al. 2014). The results from these studies

53 demonstrate a great potential of this approach, which warrants further investigation.

54   The research presented in this paper compared the use of seven individual machine

55 learning models, including M5Prime (M5P), REPTree, M5Rules, decision stump, multilayer

56 perceptron, SMO regression (SMOreg), and Gaussian processes, in predicting the compressive

57 strength of environmentally friendly concrete. It also tested two commonly used ensemble

58 methods (additive regression and bagging) by adopting each of the seven individual models as

59 the base classifier to explore the possibility of improving prediction accuracy. The ultimate goal

60 was to promote the use of data mining techniques for determining the compressive strength or

61 other properties of new types of concrete while reducing the need for extensive experiments.

62 This shift will not only save time and money for the industry, but also facilitate the use of new

63 materials. The unique set of seven data mining models was selected for exploring the prediction

64 performance of four regression tree models against other three more advanced models. This also

65 seemed to be the first time that Gaussian processes regression was examined for predicting

66 concrete strength. This research used four performance measures, namely correlation coefficient

67 (R), coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error

68 (MAE), to assess prediction accuracy of generated models. R^2 was used to compare models

69 examined in this research and previous studies.

70   This paper first introduces the unique type of environmentally friendly concrete studied in

4

71 this research and then reviews previous research efforts in modeling and predicting compressive

72 strength of concrete. A brief description of all the data mining models examined in this research

73 is presented. After describing the research methodology and experimental settings, this paper

74 presents the results and analysis as well as the findings of this research.

75 **Literature Review**

76 *Environmentally Friendly Concrete*

77 Conventional concrete is made from four main ingredients: water, cement, fine aggregate (sand),

78 and coarse aggregate. With the wide use of concrete as a building material, its negative

79 environmental impacts are significant. Specifically, the cement industry produces up to 5% of

80 global man-made CO_2 emissions (WBCSD 2009) and accounts for approximately 12–15% of

81 total industrial energy use in various countries (Madlool 2013). The concept of producing more

82 environmentally friendly concrete emerged as a response to reducing the emissions and other

83 environmental impacts from concrete production. For this purpose or as a result of a demand for

84 specific properties needed for concrete applications, alternative materials (particularly

85 supplementary cementitious materials [SCMs] and alternative aggregates) are added or used to

86 replace certain amounts of the traditional ingredients.

87   For environmentally friendly concrete, the commonly used alternative materials are those

88 that contain recycled contents, are locally available with low transportation costs, have reduced

89 greenhouse gas emissions in their production, reserve natural resources, or improve concrete

90 performance during its life cycle. Some of the frequently studied alternative aggregates include

91 recycled concrete aggregate (Etxeberria et al. 2007; Limbachiya et al. 2012), building rubbles

92 (Khalaf and Devenny 2004), fiber scrap aggregate (Shahria Alamet al. 2013), recycled glass

93 aggregate (Berry et al., 2011), etc. Fly ash (FA) class C and F (Basri et al. 1999; Kevern et al.

94   2011), furnace slag (Lubeck et al. 2012), and silica fume (Limbachiya et al. 2012), are examples

95   of materials that have been examined as SCMs.

96   For this research, Portland limestone cement (PLC), Haydite® lightweight aggregate

97   (LWA), and FA Class F were selected as alternatives to the traditional ingredients. This was

98   based on the literature review and the results of a survey that was performed by the research

99   team to identify industry interests in using environmentally friendly concrete and ingredients (Jin

100   2013). A brief review of these alternative materials can be found in Omran et al. (2014).

101   *Related Work in Modeling and Predicting Concrete Properties*

102   The experimental determination of the compressive strength of concrete, especially for concrete

103   containing alternative materials, is known to be time consuming and costly. On the other hand,

104   using simple linear regression models for prediction has limited accuracy and flexibility (Yeh

105   1998; Deepa et al. 2010). As a result, recent years have seen an increasing interest in using more

106   advanced data mining techniques for predicting concrete properties.

107   Artificial neural network (ANN) has been used to predict fresh and hardened properties of

108   high performance concrete (Khan et al. 2013) and LWA concrete (Alshihri et al. 2009; Abdeen

109   and Hodhod 2010). The results of these studies have generally confirmed ANN to be a powerful

110   method for such applications. Another widely used data mining method, Support Vector

111   Machines (SVM), has also been used to predict properties of hardened concrete, such as

112   compressive strength, tensile strength, and elastic modulus (Gupta 2007; Yan et al. 2013; Yazdi

113   et al. 2013; Aiyer et al. 2014; Akande et al. 2014). In other attempts, both ANN and SVM have

114   been applied in conjunction with fuzzy logic to improve the accuracy and reliability of prediction

115   (Nataraja et al. 2006; Saridemir et al. 2009; Cheng et al. 2012). In addition, some other

116   prediction models, e.g., ensembles of decision trees in Erdal et al. (2013), were examined for

117 predicting the compressive strength of different types of concrete. While these studies have led

118 to more accurate predictions compared to traditional regression techniques, more reliable,

119 applicable, and practicable models are yet to be discovered (Chou et al. 2011).

120   A comparison between multivariable regression analysis and ANN made by Atici (2011)

121 identified the effectiveness of these methods for predicting the strength of mineral admixture

122 concrete. With the increasing use of advanced data mining techniques in concrete property

123 prediction, a few other comparison studies were conducted to evaluate the performance of

124 multiple data mining models, mostly focused on the compressive strength prediction of high

125 performance concrete. For example, Deepa et al. (2010) examined ANN, linear regression, and

126 M5P tree model for their accuracy and time performance. Similarly, Chou et al. (2011) evaluated

127 ANN, SVM, multiple regression, multiple additive regression trees, and bagging regression trees.

128 So far, very few studies have compared multiple data mining methods in predicting the

129 compressive strength of environmentally friendly concrete. This paper aims to fill this gap and

130 provide a more accurate and reliable tool to predict the compressive strength of a unique type of

131 environmentally friendly concrete made with PLC, Haydite LWA, and FA.

132 **Predictive Data Mining Techniques Examined in This Research**

133 The research was performed in two steps: 1) Examining the prediction accuracy of seven

134 individual data mining models, including the four common regression tree models (M5P,

135 REPTree, M5-Rules, and decision stump) and three more advanced predictive models

136 (multilayer perceptron, SMOreg, and Gaussian processes regression), and 2) Examining the

137 prediction accuracy of two commonly used ensemble methods (additive regression and bagging),

138 in which each of the aforementioned models was used as base classifier to evaluate the effects of

139 boosting. Kotsiantis et al. (2006) defined three mechanisms for the ensemble of regression

140   models: 1) *using a single machine learning model with different subsets of training data*, 2)

141   *using a single learning method with different training parameters*, and 3) *using different machine*

142   *learning methods*. The second step of this research adopted the first two mechanisms by using a

143   single machine learning model as base classifier for the ensemble models. Studying multiple

144   classifiers for the ensemble models can be a subject for future research. A brief review of these

145   data mining models and selected parameters is presented below.

146   ***Regression Tree Models***

147   Regression tree models have long been used in data mining as a supervised learning technique,

148   and have been widely applied to numeric prediction. Compared to some of the state-of-the-art

149   models, regression tree models may have lower prediction accuracy, but usually perform faster

150   and are easier to interpret. This research examined four commonly used regression tree models

151   as described below.

152   **M5P** is a reconstruction of the M5 algorithm introduced by Quinlan (1992) for generating

153   a tree of regression models from empirical data (Wang and Witten 1997). In a M5P model, at

154   each branch the tree stores a linear regression model that predicts the class values of the portion

155   of dataset that reaches the leaf. The dataset splits into different portions according to certain

156   attributes of the data. Standard deviation (SD) is usually used as a criterion that determines

157   which attribute is the best for splitting the dataset at each node. The attribute to be chosen is the

158   one that has the maximum expectation to reduce error, see Eq. (1):

$$EX_{error} = SD(T) - \sum \frac{|Ti|}{|T|} \times SD\ (T_i) \tag{1}$$

159
160   where T_i denotes the subset of cases that have the *i*th outcome of the potential test. The process

161   stops when a very small change happens in class values or only a few instances remain. The tree

162   will then be pruned back and a smoothing process will be performed in the end to compensate

163 sharp discontinuities between adjacent linear models (Quinlan 1992).

164 **REPTree (Reduced Error Pruning Tree)** is a fast decision tree learner that builds a

165 decision/regression tree by using information gain or variance as decision features for splitting

166 the data at the nodes. Then the generated regression tree is pruned back using the reduced-error

167 with back over-fitting technique (Witten and Frank 2005). In the context of decision trees, the

168 term "information gain" is usually equivalent to expectation value of the Kullback–Leibler

169 divergence of a conditional probability distribution (Garcia et al. 2002). For numeric attributes,

170 REPTree sorts the values once at the start of the run, and then uses the sorted list to calculate the

171 right splits in each tree node.

172 **M5-Rules** is an algorithm that uses divide-and-conquer to generate decision lists (ordered

173 sets of if-then rule) for regression problems. Holmes et al. (1999) used decision lists to make a

174 more compact and understandable model tree compared to previous models. Decision lists can

175 work with both continuous and nominal variables. M5-Rules uses the M5 algorithm to build a

176 model tree, makes a rule from the best leaf, and then works on other instances that are left in the

177 dataset according to the generated rule.

178 **Decision Stump** is a machine learning model that only consists of one-level decision tree.

179 It has one internal node (called root node), which is immediately connected to nodes in branches

180 (referred to as terminal nodes). Decision stump makes a prediction based on the value of just a

181 single input attribute. It performs regression based on the mean squared error where each root

182 node represents an attribute in an instance to be evaluated, and each branch represents a value

183 that the node can take (Iba and Langley 1992). Decision stump is usually used as a component

184 of a boosting algorithm to improve its prediction accuracy.

185 *Multilayer Perceptron (ANN)*

186   ANN is a computational system consisting of simple, highly interconnected processing elements

187   (nodes or neurons) that work together to solve specific problems (Caudill 1987). It is an

188   algorithm inspired by research in biological nervous systems to generate a simplified model of

189   how the brain works (Rumelhart et al. 1994). The first neural network was proposed by

190   McCulloch and Pitts (1943), and since then many other models have been introduced. The basic

191   structure of a multilayer perceptron ANN model is shown in Fig. 1 below.

192   ANN models usually consist of an input layer, one or more hidden layers, and an output

193   layer. Each of these layers can have different number of nodes. Each node under the hidden

194   layer(s) will receive one or more inputs. The inputs will be multiplied by their weights, and

195   summed together and with the bias (threshold). The weighting and bias values will be initially

196   chosen as random numbers and will then be adjusted according to the results of the training

197   process (Atici 2011). The output of each node will be generated based on the significance of the

198   summation value and by the means of a predefined specific activation function, e.g., unipolar

199   sigmoid function, bipolar sigmoid function, hyperbolic tangent function, etc. (Bishop 2006).

200   *SMOreg-based SVM*

201   SVM is a supervised learning model developed by Cortes and Vapnik (1995). It has been

202   intensively used in many data mining problems for both classification and regression purposes.

203   In an SVM algorithm, the training set is first mapped to an n-dimensional feature space by using

204   a nonlinear kernel mapping procedure. Then a hyperplane, a subspace that is one dimension less

205   than its surrounding space, will be identified in this feature space according to the projected

206   dataset. The aim is to find the optimal hyperplane that separates the data points in the classes,

207   while simultaneously maximizing the margin (i.e., the distance between the hyperplane and the

208   closest points of the training set) for linearly separable patterns (Leskovec et al. 2014). The

209 hyperplane $f(x, w)$ is represented by a linear function in the feature space:

$$f(x, w) = \sum_{j=1}^{m} w_j \, g_j(x) + b \tag{2}$$

210 where $g_j(x)_{,j=1,\dots,m}$ denotes a set of nonlinear transformations, and b is the "bias" term. For

211 SVM regression purposes, Cortes and Vapnik (1995) suggested to use a so called ε, the

212 insensitive loss function that penalizes error only if it is greater than ε (Shevade et al. 2000). So

213 the $|\xi|_\varepsilon$ is represented as:

$$|\xi|_\varepsilon = \begin{cases} 0 & if \ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{3}$$

214
215   Using (non-negative) slack variables ξ_i and ξ_i^*, the final optimization problem to be solved

216 can be formulated as:

$$\text{Minimize} \quad \frac{1}{2}||w||^2 + C \sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

217   Subjected to:

$$\begin{cases} y_i - f(x_i, w) \leq \varepsilon - \xi_i^* \\ f(x_i, w) - y_i \leq \varepsilon - \xi_i \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n \end{cases} \tag{5}$$

218
219 SVM regression finds the linear regression in the high-dimension feature space using ε while

220 reducing the model complexity by minimizing $||w||^2$.

221   Sequential minimal optimization (SMO), an algorithm introduced by Platt (1998), is used

222 to solve the very large quadratic programming (QP) optimization problems in SVM through

223 breaking them into a series of smallest possible QP problems. In this way problems can be

224 solved analytically, eliminating the need for numerical optimization algorithms (Platt 1998).

225 ***Gaussian Processes***

226 Gaussian process is a powerful non-linear prediction tool, which can be used for Bayesian

11

227 regression as well as in learning process of both supervised and unsupervised learning

228 frameworks (Bishop 2006). It is a non-parametric stochastic process that generalizes the

229 Gaussian probability distribution. A Gaussian process sometimes is described as a distribution

230 over functions ($P(f)$), where f is a function that projects input space (vector **x**) to feature space

231 (vector **r**) and for any finite subset of X the marginal distribution over that subset $P(f)$ has a

232 Gaussian distribution. The f could be an infinite-dimensional quantity. As a result, Gaussian

233 process extends multivariate Gaussian distributions to infinite dimensionality (Rasmussen and

234 Williams 2006). Same as Gaussian distribution that can be specified by a mean vector $\boldsymbol{\mu}$ and

235 covariance matrix $\sum$, a Gaussian process can be defined by a mean function $M(x)$ and

236 covariance function $k(x, x')$ expressed as $f(x) \sim GP\big(M(x), k(x, x')\big)$. One of the advantages of

237 a Gaussian process model is that its formulation is probabilistic. This is especially useful for

238 probabilistic prediction and "gives the ability to infer model parameters such as those that control

239 the kernel shape and the noise level" (Chu and Ghahramani 2006).

240 **Ensemble Methods Used in This Research**

241 According to Rokach (2010), the idea of ensemble learning models started with Tukey (1977) at

242 late 1970s by simply combining two linear regression models using residual of the first model

243 for the second modeling process. This effort was then followed by many other attempts, such as

244 partitioning the input space and using two or more classifiers (Dasarathy and Sheela 1979) or

245 using the AdaBoost algorithm (Freund and Schapire 1996). The purpose for ensemble modeling

246 is to achieve better prediction performance by combining multiple learning algorithms.

247 *Additive Regression (Gradient Boosting)*

248 Regression trees are well known for many advantages such as flexibility of input variables (e.g.,

249 numeric, ordinal, binary, and categorical variables) and immunity to the effects of extreme

250   outliers. However, these methods usually suffer from the lack of accuracy. Gradient boosting,

251   first introduced by Friedman (2001), is an additive regression tree model that can overcome this

252   drawback through the application of a boosting technique (Friedman and Meulman 2003).

253   Additive models are nonparametric regression methods, which assume that each input feature

254   has a separate contribution to the final prediction and these input features can be added up to

255   generate the regression model for prediction (Friedman and Stuetzle 1981). According to

256   Friedman and Meulman (2003), boosting a tree-based model can significantly increase its

257   prediction accuracy. Additive regression is a metadata learner that improves the performance of

258   weak prediction models (e.g., regression tree models) by applying the stochastic gradient

259   boosting technique. The technique mainly involves fitting sequence of models: The first model in

260   the sequence is trained based on the original dataset, and each of the next models will be trained

261   on a new dataset containing the residual errors remained from fitting the previous model

262   (Friedman, 2001).

263   ***Bagging***

264   Bagging is short for Bootstrap Aggregating. Breiman (1994) defines bagging as a way to

265   generate multiple versions of a predictor, through which a more robust predictor can be

266   generated. It is an ensemble meta-algorithm that improves the accuracy and stability of the

267   prediction. The algorithm is based on generating bootstrap replications of dataset and using these

268   different versions of dataset as new training sets to generate multiple models. The final

269   prediction is achieved through combining the outcomes of these models (i.e., averaging the

270   results for the regression problem and using plurality voting for the classification problem).

271   Previous studies have shown that bagging can significantly improve the results of unstable

272   models (e.g., models sensitive to small changes in the training dataset), models with high

13

273   dimensional dataset problems, and classification and regression tree models (Breiman 1994;

274   Buhlmann and Yu 2002).

275   **Methodology and Experimental Settings**

276   *Concrete Experimental Design and Data Collection*

277   In this study, 36 different batches of concrete were designed and prepared. Each batch contained

278   different replacement percentages of fly ash Class F (0%, 20%, 30% or 40%) and Haydite LWA

279   (0%, 33%, 67% or 100%) besides the use of either Portland cement (PC) Type I/II or PLC Type

280   GUL. In this way, the effects of alternative materials on the compressive strength of concrete can

281   be examined more accurately. The fly ash Class F replaced part of PC or PLC by different

282   percentages of weight and Haydite LWA substituted pea gravel by different percentages of

283   volume. Their numerical values were used as inputs for the tested models. In addition to the

284   above three variables, the actual water content, the amounts of sand, pea gravel, and Micro Air®,

285   as well as the concrete curing age were selected as the other influential variables for the models.

286   Table 1 shows the range, mean, and SD of those variables in this experimental study.

287   All the concrete mixed in the experiment was assumed to be air-entrained (considered to

288   be used outdoors in cold climate) by adding Micro Air, an air entraining agent, to the mixtures.

289   The intended slump was 12.70 - 15.24 cm and the air content was 6-7%. Concrete was mixed in

290   a laboratory mixer and the whole processes of making, pouring and curing concrete were

291   performed based on ASTM C 31/C 31 M – 06 guideline. Three 10.16 cm by 20.32 cm cylinders

292   from each batch of concrete mixture were tested in each of four different curing ages of 3, 7, 28

293   and 90 days for compressive strength. The average test result of each three cylinders formed a

294   data point in the database. All the details for the experiments can be found in Jin (2013).

295   *Parameter Setting of Data Mining Models*

296 In this study, the Weka workbench toolbox (Waikato 2015) was used to generate the examined

297 machine learning models for predicting the compressive strength of the environmentally friendly

298 concrete. Since one of the original goals for experimental testing was to compare the

299 compressive strength of PC and PLC concrete, this research performed a simple paired T-Test on

300 the PC and PLC concrete datasets, which confirmed a statistical difference between these two

301 groups. To evaluate the potential impact of the statistically different datasets on the prediction

302 accuracy of data mining models, this research took the following three-step approach: The first

303 was to test the selected data mining models based on the PC or PLC dataset only. In such cases,

304 seven variables were used to generate the models. The second step was to examine the selected

305 models based on the whole dataset including all PC and PLC concrete samples. In the modeling

306 process, eight variables including a new binary variable "cement type" were used. Thirdly, the

307 prediction performance of data mining models based on different datasets was compared to learn

308 whether simpler models with seven variables and individual datasets will lead to better

309 prediction accuracy, or the prediction accuracy can be improved by a larger sample size though

310 additional variable(s) may be needed, leading to more complex models.

311   An example of ANN model with eight input variables including "cement type" is shown in

312 Fig. 2. The symbol "8-3-1" means that there are eight nodes in the input layer, three nodes in the

313 hidden layer, and only one node at the output of the network.

314   Many input parameters need to be set up for most data mining algorithms. The setting of

315 input parameters could affect the accuracy and/or reliability of generated models. In this research,

316 a comprehensive sensitive analysis was carefully conducted on each model to identify the

317 parameter values/options that could lead to the highest prediction accuracy among all the

318 examined model settings, while avoiding over-fitting issues. The important parameters that were

319   tested in this sensitivity analysis are presented in the Analytical Results and Discussion section.

320   *Performance Measures*

321   The models were trained with different parameters and/or variables. Their prediction accuracy

322   was evaluated and compared based on four frequently used performance measurements in

323   previous studies: R, R^2, RMSE, and MAE. R, RMSE, and MAE are formulated as:

$$R = \frac{\sum_{i=1}^{n}(P_i - \mu_P)(A_i - \mu_A)}{\sqrt{\sum_{i=1}^{n}(P_i - \mu_P)^2 \ \sum_{i=1}^{n}(A_i - \mu_A)^2}} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{7}$$

$$MAE = \frac{\sum_{i=1}^{n}|P_i - A_i|}{n} \tag{8}$$

324   where A_i and P_i represent the actual and predicted compressive strength of concrete samples

325   related to data point *i*, respectively, *n* is the total number of data points in the validation set(s), μ_A

326   is the mean value of observations, and μ_P is the mean value of predictions.

327   A 10-fold cross-validation was used in this study to minimize the bias associated with the

328   random sampling of the training and holdout data samples in regular validation methods. The

329   cross-validation is a technique that evaluates the expected accuracy and validity of a predictive

330   model by dividing a dataset into different subsets and evaluating the accuracy of the model for

331   each of those subsets. In general, a k-fold cross-validation includes the following steps:

332   •   Splitting the dataset into K subsets of equal size (K folds)

333   •   In each run, training the model on all the subsets except one

334   •   Evaluating the prediction accuracy by using the left out subset to test the trained model

335   •   Repeating steps 2 and 3 for K times and each time leaving a different fold for testing

336   •   Calculating the final performance measurements by averaging the performance

16

337   measurements from each of the K runs.

338   This would improve the generalization and reliability of the performance measurements obtained

339   for models under testing.

340   **Analytical Results and Discussion**

341   *Characteristics of Concrete Datasets Used in This Research*

342   The datasets used in this research, i.e., the measured compressive strength of PC and PLC

343   concrete samples, are illustrated in Fig. 3. It can be noted that the dramatic changes in

344   compressive strength values shown in the figure were caused by different curing ages of the

345   tested samples (3, 7, 28, and 90 days). The variation of compressive strength values obtained at

346   the same curing ages was caused by different concrete mix designs.

347   Table 2 shows the results of the paired T-Test, which suggest that a statistically significant

348   difference exists between the average compressive strength of PC- and PLC-based concrete. In

349   other words, it shows that with 95% confidence, the average compressive strength of PLC

350   concrete samples is 2.76 to 4.36 MPa higher than that of PC concrete samples.

351   *Comparison Results for the Data Mining Models Tested*

352   In the following, comparison results for the data mining models tested in this research are

353   presented. Due to its poor prediction accuracy (e.g., R values at 0.5226, 0.6001 and 0.6208 for

354   the PLC, PC and combined datasets, respectively), the decision stump model is excluded from

355   most of the tables and figures presented below. The exception is for the presentation of results

356   related to ensemble models. This is because this study found that when decision stump was used

357   as the base classifier for the ensemble models, the prediction accuracy was acceptable.

358   Fig. 4 shows the highest R achieved by each of the eight data mining models. It was found

359   that the prediction accuracy increased in five of the tested models when combining the two

17

360   datasets (PC and PLC) and using the cement type as an additional binary input. Exceptions are

361   the three regression tree models (i.e., M5P, REPTree and M5-Rules), in which the accuracy of

362   prediction based on the PC concrete dataset was slightly better than the combined dataset.

363   The measured performance of prediction models in terms of RMSE and MAE is presented

364   in Tables 3 and 4. The bolded value in each row represents the highest prediction accuracy

365   achieved in this study when different datasets were used for testing individual and ensemble

366   models. It seems that according to both criteria (MAE and RMSE), additive regression based on

367   the Gaussian processes classifier obtained the highest prediction accuracy for comprehensive

368   strength of PLC samples while the individual Gaussian processes regression model achieved the

369   highest prediction accuracy for both the PC and combination datasets.

370   The information presented above shows that the listed models all had acceptable prediction

371   performance. Further, the Gaussian processes regression model achieved the best prediction

372   accuracy based on all the three performance measures while REPTree had the lowest. Table 5

373   below lists the important parameters and associated values/options used for these models to

374   achieve their highest prediction accuracy. In particular, the option of "polykernel" was selected

375   for all of the four models that need a kernel as their covariance matrix. These include additive

376   regression, bagging, Gaussian processes, and SMOreg. From this point forward, the analysis and

377   results are solely presented for the combined (PC & PLC) dataset, which was proven to have

378   improved the prediction accuracy for most models tested in this study.

379   Fig. 5 illustrates the relationship between the predicted and actual compressive strength of

380   the studied concrete samples for each of the eight predictive models. All the plots show fairly

381   linear relationships between predicted and actual values. Apparently, the Gaussian processes

382   regression model is the best representative of actual experimental data with the highest R^2 at

18

383   0.9842.

384   Fig. 6 displays the distribution of residuals and percentage error for the tested models. It is

385   observed that in all these plots when the actual compressive strength of concrete samples

386   increased, residuals became larger but the associated percentage errors decreased. Similar to the

387   early findings, Gaussian processes regression, bagging, and additive regressions are the models

388   with prediction results being the closest to the actual experimental values.

389   Table 6 compares R values achieved by the seven individual data mining models as well as

390   two ensemble methods with each of individual data mining models used as base classifier. The

391   comparison results show that both the additive regression and bagging algorithms using

392   regression tree models as base classifier achieved better prediction accuracy than individual

393   regression tree models. On the other hand, when SMOreg, Gaussian processes, and multilayer

394   perceptron were used as base classifier, mixed results were generated. Similar to the early

395   conclusion from the individual model comparison, the highest accuracy of prediction for additive

396   regression and bagging was all achieved when the Gaussian processes was used as their base

397   classification model. This finding is particularly important since Gaussian processes regression

398   has rarely been applied in existing research to predict concrete properties.

399   Table 7 lists the average time spent for building each of the tested models. These times

400   were associated with the parameter settings for these models to achieve the highest prediction

401   accuracy in the sensitivity analysis. Due to the use of 10-fold cross validation, the training time

402   for each of these models was much longer than the time used to build the model. Although many

403   variables could affect the length of the training time, the total time was mostly proportional to

404   the time used to build the model. The results indicate that even though the three more advanced

405   predictive models achieved higher prediction accuracy in general they are far more time-

19

406 consuming compared to individual regression tree models as well as ensemble models with

407 regression tree as base classifier. The individual Gaussian processes model was somewhat an

408 exception with relatively fast building and training time.

409 *Comparison with Previous Work*

410 Table 8 provides a brief comparison of the highest prediction performance achieved in this study

411 and some of the primary previous works that used data mining models to predict the compressive

412 strength of concrete. The comparison of R^2 values obtained by different studies shows that eight

413 of the data mining models examined in this research offered fairly high prediction accuracy with

414 R^2 ranging from 0.9217 to 0.9842. Moreover, compared with the same types of models

415 examined in previous research, i.e., M5P, SVM, bagging, and additive regression, this study

416 achieved relatively better prediction performance. It is worth noting that this research applied the

417 cross validation method for evaluating the accuracy of predictions, which was not the case in

418 most of previous studies listed in Table 9 except for Chou et al. (2011) and Deepa et al. (2010).

419 Compared to the traditional validation method, cross validation usually lowers the R^2 values of

420 tested models, but improves the generalization and reliability of the assessment.

421 According to Table 8, the Gaussian processes regression model provided the highest

422 prediction accuracy ($R^2 = 0.9837$) among all the data mining models compared, while having a

423 relatively fast modeling speed. Based on the extent of literature review performed by the authors,

424 this research seemed to be the first work that examined Gaussian processes regression for

425 predicting concrete properties, suggesting a great need for future research. Further, in most cases,

426 ANN led to higher prediction accuracy than traditional modeling approaches such as linear

427 regression or regression tree models.

428 In this research, the additive regression model would rank first in prediction accuracy when

429   without the presence of Gaussian processes regression, which is consistent with the results from

430   Chou et al. (2011). However, Chou et al. used decision stump as base classifier; this research

431   found that additive regression based on decision stump had the lowest accuracy and the other six

432   tested base classifiers could improve the prediction performance of additive regression. Also, in

433   Chou et al. (2011), the prediction performance of bagging with the base fast decision tree leaner

434   was not as good as the ANN model. In contrast, this study found that bagging could provide

435   better prediction accuracy than the ANN model when using the advanced methods (i.e., Gaussian

436   processes regression and multilayer perceptron) as base classifiers.

437   **Conclusions**

438   This research aimed to evaluate the potential of using data mining techniques for predicting the

439   compressive strength of environmentally friendly concrete containing fly ash, Haydite LWA,

440   and/or PLC. In particular, four common regression tree models (M5P, REPTree, M5-Rules, and

441   decision stump) and three more advanced predictive models (ANN based on multilayer

442   perceptron, SMOreg-based SVM regression, and Gaussian processes regression) were generated

443   and tested individually. Then they were used as base classifiers in two ensemble models

444   (additive regression and bagging) to evaluate the effects of boosting.

445   The obtained analytical results suggest that all of the tested models, except for decision

446   stump, can provide acceptable prediction accuracy with R^2 ranging from 0.9217 (for REPTree) to

447   0.9842 (for Gaussian processes regression). The Gaussian processes regression model showed

448   the best prediction accuracy as an individual data mining model. Also, when used as base

449   classifier, it helped the two ensemble models achieve the best prediction performance. This

450   observation is important since the Gaussian processes regression model is rarely investigated in

451   previous works in this field.

452   The results of this research also indicate that in most cases, except for M5P, REPTree, and

453   M5-Rules, training the models with the combined dataset containing PC and PLC concrete

454   samples provided better prediction accuracy than using only the PC or PLC dataset. Furthermore,

455   although the prediction accuracy of the three advanced data mining models was higher than that

456   of the four regression tree models, the time required for building and training the models was

457   significantly longer. This should be considered a factor in choosing an appropriate data mining

458   model in practice. Particularly, when dealing with a very large dataset, using an ensemble

459   method with a regression tree base classifier seems to be a more practical alternative. With the

460   demonstrated potential of using data mining models to predict concrete comprehensive strength,

461   future research can adopt this approach to study other properties of concrete such as tensile

462   strength, durability, or concrete slump.

463 **References:**

464 Abdeen, M. A. M., and Hodhod, H. (2010). "Experimental investigation and development of

465   artificial neural network model for the properties of locally produced light weight aggregate

466   concrete." *Sci. Res. Org. Eng.*, 2(6), 408-419.

467 Aiyer, B. G., Kim, D., Karingattikkal, N., Samui, P., and Rao, P. R. (2014). "Prediction of

468   compressive strength of self-compacting concrete using least square support vector machine

469   and relevance vector machine." *KSCE J. Civ. Eng.,* 18(6), 1753-1758.

470 Akande, O. K., Owolabi, O. T., Twaha, S., and Olatunji, S. O. (2014). "Performance comparison

471   of SVM and ANN in predicting compressive strength of concrete." *IOSR J. Comput.*

472   *Eng.*, 16(5), 88-94.

473 Alshihri, M. M., Azmy, A. M., and El-Bisy, M. S. (2009). "Neural networks for predicting

474   compressive strength of structural light weight concrete." *Constr. Build. Mater.*, 23(6), 2214-

475   2219.

476 Atici, U. (2011). "Prediction of the strength of mineral admixture concrete using multivariable

477   regression analysis and an artificial neural network." *Expert Syst. Appl.,* 38(8), 9609-9618.

478 Basri, H. B., Mannan, M. A., and Zain, M. F. M. (1999). "Concrete using waste oil palm shells

479   as aggregate." *Cem. Concr. Res.*, 29(4), 619–622.

480 Berry, M., Stephens, J., and Cross, D. (2011). "Performance of 100% fly ash concrete with

481   recycled glass aggregate." *ACI Mater. J.,* 108 (4), 378-384.

482 Bishop, C. M. (2006). *Pattern recognition and machine learning*, Springer, New York.

483 Buhlmann, P., and Yu, B. (2002). "Analyzing bagging." *Ann. Statist.,* 30, 927-961.

484 Breiman, L. (1994). "Bagging predictors." *Mach. Learn.,* 24(2), 123-140.

485 Caudill, M. (1987). "Neural networks primer." *J. AI Expert*, 2(12), 46-52.

486   Cheng, M. Y., Chou, J. S., Roy, A. F., and Wu, Y. W. (2012). "High-performance concrete

487   compressive strength prediction using time-weighted evolutionary fuzzy support vector

488   machines inference model." *Autom. Constr.,* 28, 106-115.

489   Chou, J. S., Chiu C. K., Farfoura M., and Al-Taharwa I. (2011). "Optimizing the prediction

490   accuracy of concrete compressive strength based on a comparison of data-mining

491   techniques." *J. Comput. Civ. Eng.,* 25(3), 242-253.

492   Chu, W., and Ghahramani, Z. (2006). "Gaussian Processes for Ordinal Regression." *J. Mach.*

493   *Learn. Res.,* 6(7), 1019-1042.

494   Cortes, C., and Vapnik, V. (1995). "Support-vector networks." *Mach. Learn.,* 20(3), 273-297.

495   Dasarathy, B. V., and Sheela, B. V. (1979). "A composite classifier system design: Concepts

496   and methodology." *Proc. IEEE*., 67(5), 708-713.

497   Deepa, C., Sathiyakumari, K., and Sudha, V. (2010). "Prediction of the compressive strength of

498   high performance concrete mix using tree based modeling." *Int. J. Comput. Appl. T.,* 6(5),

499   18-24.

500   Erdal, H. I., Karakurt, O., and Namli, E. (2013). "High performance concrete compressive

501   strength forecasting using ensemble models based on discrete wavelet transform." *Eng. App.*

502   *Artif. Intell.,* 26(4), 1246-1254.

503   Etxeberria, M., Vázquez, E., Marí, A., and Barra, M. (2007). "Influence of amount of recycled

504   coarse aggregates and production process on properties of recycled aggregate concrete." *Cem.*

505   *Concr. Res.*, 37(5), 735-742.

506   Fazel Zarandi, M. H., Türksen, I. B., Sobhani, J., and Ramezanianpour, A. A. (2008). "Fuzzy

507   polynomial neural networks for approximation of the compressive strength of concrete."

508   *Appl. Soft Comput.,* 8(1), 488–498.

509 Freund, Y., Schapire, R. E. (1996). "Experiments with a new boosting algorithm." *Mach. Learn.:*

510 *Proc., 13ᵗʰ Int. Conf.*, Bari, Italy, 148-156.

511 Friedman, J. H. (2001)."Greedy function approximation: A gradient boosting machine." *Ann.*

512 *Statist.,* 29(5), 1189-1232.

513 Friedman, J. H., and Meulman, J. J. (2003). "Multiple additive regression trees with application

514 in epidemiology." *Stat. Med.,* 22(9), 1365-1381.

515 Friedman, J. H., and Stuetzle, W. (1981). "Projection pursuit regression." *J. Amer. Statist.*

516 *Assoc.,* 76(376), 817-823.

517 Garcia, J., Fdez-Valdivia, J., Rodriguez-Sanchez, R., and Fdez-Vidal, X. (2002). "Performance

518 of the Kullback-Leibler information gain for predicting image fidelity." *Proc., 16ᵗʰ Int. Conf.*

519 *Patt. Recog.,* Vol. 3, IEEE Computer Society Press, Washington, D.C, 843-848.

520 Gupta, R., Kewalramani, M. A., and Goel, A. (2006). "Prediction of concrete strength using

521 neural-expert system." *J. Mater. Civ. Eng.*, 18(3), 462–466.

522 Gupta, S. M. (2007). "Support vector machines based modelling of concrete strength." *World*

523 *Acad. Sci.: Eng. Technol.,* 3(1), 12-18.

524 Holmes, G., Hall, M., and Frank, E. (1999). "Generating rule sets from model trees." *Lect. Notes*

525 *Comput. Sci.,* 1747, 1-12.

526 Iba, W., and Langley, P. (1992)."Induction of one-level decision trees." *Proc., 9ᵗʰ Int. Conf.*

527 *Mach. Learn.,* Morgan Kaufmann, San Francisco, CA, 233–240.

528 Jin, R. (2013). *A statistical modeling approach to studying the effects of alternative and waste*

529 *materials on green concrete properties*, Ph.D. Dissertation, The Ohio State University,

530 Columbus, OH. <http://rave.ohiolink.edu/etdc/view.cgi?acc%5Fnum=osu1372854071>.

531 Kevern, J. T., Schaefer, V. R., and Wang, K. (2011). "Mixture proportion development and

532 performance evaluation of pervious concrete for overlay applications." *ACI Mater. J.,* 108(4),

533 439-448.

534 Khalaf, F. M., and Devenny, A. S. (2004). "Recycling of demolished masonry rubbles as coarse

535 aggregate in concrete: Review." *J. Mater. Civ. Eng.*, 16(4), 331–340.

536 Khan, S. U., Ayub, T. F. A., and Rafeeqi, S. (2013). "Prediction of compressive strength of plain

537 concrete confined with ferrocement using artificial neural network (ANN) and comparison

538 with existing mathematical models." *Amer. J. Civ. Eng. Arch.*, 1(1), 7-14.

539 Kotsiantis, S., Kanellopoulos, D., and Zaharakis, I. (2006). "Bagged averaging of regression

540 models." *Int. Feder. Inf. Process. Publications (Ifip),* 204, 53-60.

541 Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets,* Cambridge

542 University Press, U.K. <http://proquest.safaribooksonline.com/9781316147047> (Jun. 6,

543 2015).

544 Limbachiya, M., Meddah, M. S., and Ouchagour, Y. (2012). "Performance of Portland/silica

545 fume cement concrete produced with recycled concrete aggregate." *ACI Mater. J.,* 109(1),

546 91-100.

547 Lubeck, A., Gastaldini, A., Barin, D., and Siqueira, H. (2012). "Compressive strength and

548 electrical properties of concrete with white Portland cement and blast-furnace slag." *Cem.*

549 *Concr. Compos.,* 34(3), 392-399.

550 Madlool, N., Saidur, R., Rahim, N., and Kamalisarvestani, M. (2013). "An overview of energy

551 savings measures for cement industries." *Renew. Sust. Energ. Rev.*, 19, 18-29.

552 McCulloch, W. S., and Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous

553 activity." *Bull. Math. Biophys.*, 5(4), 115-133.

554 Nataraja, M. C., Jayaram, M. A., and Ravikumar, C. N. (2006)."A Fuzzy-Neuro model for

555   normal concrete mix design." *Eng. Letters*, 13(3), 98-107.

556   Omran, B. A., Chen, Q., and Jin, R. (2014). "Prediction of compressive strength of "green"

557   concrete using artificial neural networks." *Proc., 50th ASC Ann. Int. Conf.*, Associated

558   Schools of Construction (ASC), Windsor, CO.

559   Platt, J. C. (1998). *Sequential minimal optimizer: A fast algorithm for training support vector*

560   *machines*, Technical Report MSR-TR-98-14, Microsoft Research, Redmond, WA.

561   Quinlan, J. R. (1992)."Learning with continuous classes." *Proc., 5th Australian Joint Conf. Artif.*

562   *Intell.*, Sydney, Australia, 343–348.

563   Rasmussen, C. E., and Williams, C. K. I. (2006). *Gaussian processes for machine learning*, MIT

564   Press, Cambridge, MA.

565   Rokach, L. (2010). "Ensemble-based classifiers." *Artif. Intell. Rev.: Int. Sci. Eng. J.*, 33(1-2), 1-

566   39.

567   Rumelhart, D. E., Widrow, B., and Lehr, M. A. (1994). "The basic ideas in neural networks."

568   *Commun. ACM*, 37(3), 87-92.

569   Saridemir M., Ozcan F., Severcan M. H., and Topçu I. B. (2009). "Prediction of long-term

570   effects of GGBFS on compressive strength of concrete by artificial neural networks and

571   fuzzy logic." *Constr. Build. Mater.*, 23(3), 1279-1286.

572   Shahria Alam, M., Slater, E., and Muntasir billah, A. H. M. (2013). "Green concrete made with

573   RCA and FRP scrap aggregate: fresh and hardened properties." *J. Mater. Civ. Eng.,* 25(12),

574   1783-1794.

575   Shevade S. K., Keerthi S. S., Bhattacharyya C, and Murthy, k. R. k. (2000). "Improvements to

576   the SMO algorithm for SVM regression." *IEEE Trans. Neural Net.,* 11(5), 1188-93.

577   Topçu, I. B., and Saridemir, M. (2007). "Prediction of properties of waste AAC aggregate

578   concrete using artificial neural network." *Comp. Mater. Sci.,* 41(1), 117-125.

579   Tukey, J. W. (1977). *Exploratory data analysis*, Addison-Wesley Pub. Co., Reading, MA.

580   Vilalta, R., and Drissi, Y. (2002). "A perspective view and survey of meta-learning." *Artif. Intell.*

581   *Rev.,* 18(2), 77-95.

582   Vilalta, R., Giraud-Carrier, C., Brazdil, P., and Soares, C. (2004)."Using meta-learning to

583   support data-mining." *Int. J. Comput. Sci. Appl.*, 1(1), 31-45.

584   Wang, Y., and Witten, I. H. (1997). "Introduction of model trees for predicting continuous

585   classes." *Proc., 1997 Euro. Conf. Mach. Learn.*, University of Economics, Faculty of

586   Informatics and Statistics, Prague.

587   Witten, I. H., and Frank, E. (2005). *Data mining: Practical machine learning tools and*

588   *techniques*, 2nd Ed., Morgan Kaufman, San Francisco, CA.

589   Waikato. (2015). <http://www.cs.waikato.ac.nz/ml/index.html> (Jun. 6, 2015).

590   Yang, E. I., Yi, S. T., and Leem, Y. M. (2005). "Effect of oyster shell substituted for fine

591   aggregate on concrete characteristics: Part I. Fundamental properties." *Cem. Concr. Res.*,

592   35(11), 2175-2182.

593   Yan, K., Xu, H., Shen G., and Liu P. (2013). "Prediction of splitting tensile strength from

594   cylinder compressive strength of concrete by support vector machine." *Adv. Mater. Sci. Eng.,*

595   10.1155/2013/597257.

596   Yazdi, J. S., Kalantary, F., and Yazdi, H. S. (2013). "Prediction of elastic modulus of concrete

597   using support vector committee method." *J. Mater. Civ. Eng.*, 25(1), 9-20.

598   Yeh, I. C. (1998). "Modeling of strength of high performance concrete using artificial neural

599   networks." *Cem. Concr. Res.,* 28(12), 1797–1808.

600   Yeh, I. C., and Lien, L.-C. (2009). "Knowledge discovery of concrete material using genetic

601   operation trees." *Expert Syst. Appl*., 36(3), 5807–5812.

602   WBCSD (World Business Council for Sustainable Development). (2009). *"Cement technology*

603   *roadmap 2009 - Carbon emissions reductions up to 2050."*

604   <http://www.wbcsd.org/Pages/EDocument/EDocumentDetails.aspx?ID=11423&NoSearchC

605   ontextKey=true> (Jun. 6, 2015).

606   Zhang, G. (1998). "Forecasting with artificial neural networks: The state of the art." *Int. J.*

607   *Forecasting.,* 14(1), 35-62.

608 **Table List**

609   **Table 1.** Parameters and Values for Concrete Mix Design (Per Cubic Meter of Concrete)

610   **Table 2.** Paired T-Test of the Means for PC and PLC Concrete

611   **Table 3.** The Lowest MAE Calculated for Each of the Models based on Different Datasets

612   **Table 4.** The Lowest RMSE Calculated for Each of the Models based on Different Datasets

613   **Table 5.** Important Parameters and Associated Values/Options for Achieving the Highest

614   Accuracy of the Tested Models

615   **Table 6.** R Values for Ensemble Models Using Different Classifiers

616   **Table 7.** Time (in Second) for Building Each Model

617   **Table 8.** Comparison of Prediction Accuracy with Previous Works

618   **Figure List**

619   **Fig. 1.** Structure of ANN models

620   **Fig. 2.** The example ANN model (8-3-1) created for this research

621   **Fig. 3.** Experimental results for compressive strength of PC and PLC concrete

622   **Fig. 4.** The highest R value for each of the models based on different datasets

623   **Fig. 5.** Predicted vs. actual compressive strength (abbreviated as CS in the figure)

624   **Fig. 6.** Residuals and percentage errors vs. actual compressive strength values

625 **Table 1.** Parameters and Values for Concrete Mix Design (Per Cubic Meter of Concrete)

Parameter	Min.	Max.	Mean	SD
Age (day)	3	90	35.12	35.37
Water (kg)	210.61	210.61	210.61	0
PC or PLC (kg)	226.63	528.02	346.18	102.07
Fly ash (kg)	0	211.21	79.80	72.37
Sand (kg)	741.60	901.78	768.29	59.91
Pea gravel (kg)	0	750.49	483.40	229.54
Haydite (kg)	0	368.42	131.13	113.03
Micro Air (ml)	112.17	135.38	123.78	11.64

626

627   **Table 2.** Paired T-Test of the Means for PC and PLC Concrete

Statistical item	Compressive strength for PLC concrete (MPa)	Compressive strength for PC concrete (MPa)
Mean	37.109125	33.54779
Variance	227.65862	195.3950
Observations	72	72
Hypothesized mean difference	0	
t stat	8.8572	
p(T<=t) one-tail	2.16E-13	
t critical one-tail	1.6665	
p(T<=t) two-tail	4.31E-13	
t critical two-tail	1.9939	

628

629 **Table 3.** The Lowest MAE Calculated for Each of the Models based on Different Datasets

Method	Additive Regression	Bagging	M5P	REPTree	M5-Rules	SMOreg	Multilayer Perceptron	Gaussian Processes
PLC	**1.52**	2.1038	3.4854	4.9203	3.9587	2.4839	1.946	1.6343
PC	1.8992	1.9536	2.4113	3.0505	2.3633	2.36	2.1796	**1.8784**
PLC & PC	1.3976	1.5662	2.4536	3.3953	2.4793	2.072	1.9625	**1.3756**

630

631   **Table 4.** The Lowest RMSE Calculated for Each of the Models based on Different Datasets

Method	Additive regression	Bagging	M5P	REPTree	M5-Rules	SMOreg	Multilayer perceptron	Gaussian processes
PLC	**2.0309**	2.6724	4.7615	6.2041	5.2028	3.3491	3.1178	2.2236
PC	2.4223	2.4563	2.9852	3.8477	2.9705	2.9571	2.9439	**2.4154**
PLC & PC	1.8624	1.9902	3.3367	4.1663	3.3169	2.6104	2.5473	**1.837**

632

633 **Table 5**. Important Parameters and Associated Values/Options for Achieving the Highest
634 Accuracy of the Tested Models
635

Data mining model	The highest R	Name of parameter/option	Associated value/option
Additive regression	0.9918	Base classifier	Gaussian process
		Number (no.) of iteration	10
		Shrinkage rate	1
		Level of Gaussian noise	0.002
		Kernel of the choice	polykernel
		Exponent value	3
Bagging	0.9907	Base classifier	Gaussian process
		No. of iteration	80
		Bagging size percentage	100
		Level of Gaussian noise	0.007
		Kernel of the choice	polykernel
		Exponent value	3
M5P	0.9735	Min. no. of instances	5
M5-Rules	0.9738	Min. no. of instances	4
REPTree	0.9601	Min. total weight of instances	1
		Min. proportion of the variance	0.0001
SMOreg	0.9839	Kernel of the choice	polykernel
		Exponent value	3
Multilayer perceptron	0.9849	Node No. for first hidden layer	15
		Node No. for second hidden layer	8
		Learning Rate	0.1
		Momentum	0.25
		Training time	10000
		Validation threshold	20
Gaussian processes regression	0.9921	Kernel of the choice	polykernel
		Exponent value	3
		Level of Gaussian noise	0.0005

636

637   **Table 6.** R Values for Ensemble Models Using Different Classifiers

Method	REPTree	M5-Rules	M5P	Decision stump	SMOreg	Gaussian processes	Multilayer perceptron
Individual model	0.9601	0.9738	0.9735	0.6208	0.9839	**0.9921**	0.985
Additive regression	0.9822	0.9778	0.9917	0.9712	0.9845	**0.9918**	0.9793
Bagging	0.9701	0.9765	0.9786	0.9421	0.9823	**0.9907**	0.9899

638   Note: The bold numbers indicate the best performance result for each dataset.

639 **Table 7.** Time (in Second) for Building Each Model

Method	REPTree	M5Rules	M5P	Decision stump	SMOreg	Gaussian processes	Multilayer perceptron
Individual model	0.02	0.14	0.05	0	10.19	0.33	42.46
Additive regression	0.03	0.42	1.92	0.17	43.82	3.26	167.36
Bagging	0.28	1.09	3.71	0.03	127.02	27.89	419.42

640

641   **Table 8.** Comparison of Prediction Accuracy with Previous Works

Previous work	Sample size	Technique	R^2
Yeh 1998	727	ANN	0.914 (avg.)[a]
		Linear regression	0.574 (avg.)[a]
Gupta et al. 2006	864	Neural-fuzzy inference system	0.76
Fazel Zarandi et al. 2008	458	Fuzzy polynomial neural networks	0.8209
Yeh and Lien 2009	1196	Genetic operation trees	0.8669
		ANN	0.9338
Chou et al. 2011	1030	ANN	0.9091
		Multiple regression	0.6112
		SVM	0.8858
		Multiple Additive Regression Trees (MART)	0.9108
		Bagging Regression Trees (BRT)	0.8904
Deepa et al. 2010	300	Multilayer perceptron (ANN)	0.625
		Linear regression	0.491
		M5P model tree	0.787
Atici 2011	135	ANN	0.9801
		Multiple regression	0.899
Erdal et al. 2013	1030	ANN	0.9088
		Bagged ANN	0.9278
		Gradient Boosted ANN	0.9270
		Wavelet Bagged ANN	0.9397
		Wavelet Gradient Boosted ANN	0.9528
This paper	144	M5P model tree	0.9476
		M5-Rules	0.9482
		REPTree	0.9217
		Multilayer perceptron (ANN)	0.970
		SMOreg (SVM)	0.968
		Gaussian processes regression	0.9843
		Additive regression	0.9837
		Bagging	0.9816

642   [a]In Yeh (1998), the database was divided into four different sets. Each time one set was used for testing and the
643   other three sets were used for training. The listed R^2 value is the average for the four testing datasets.
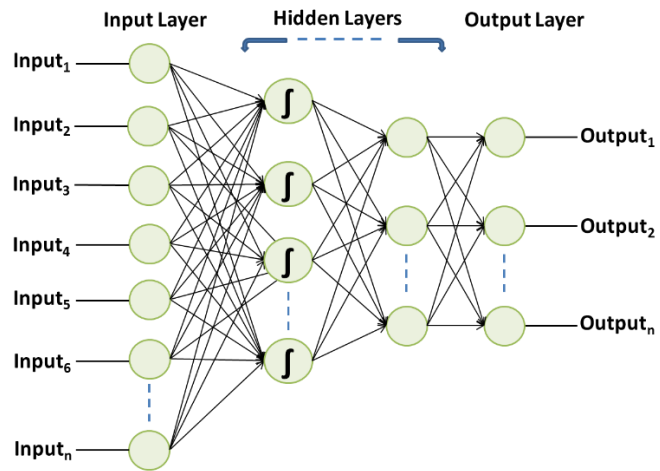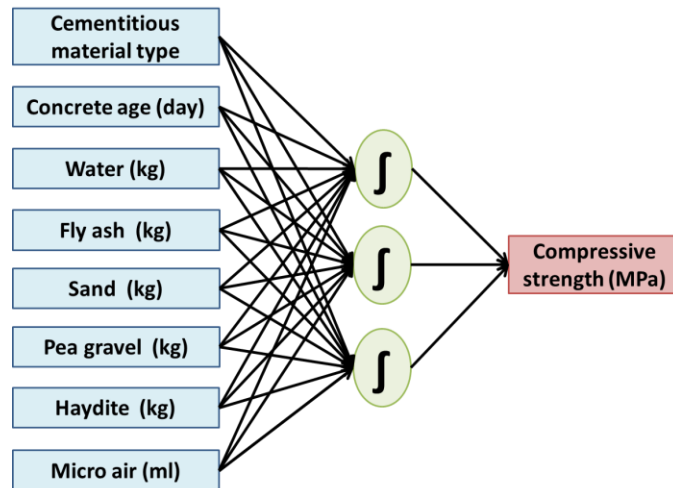
Fig. 1. Structure of ANN models
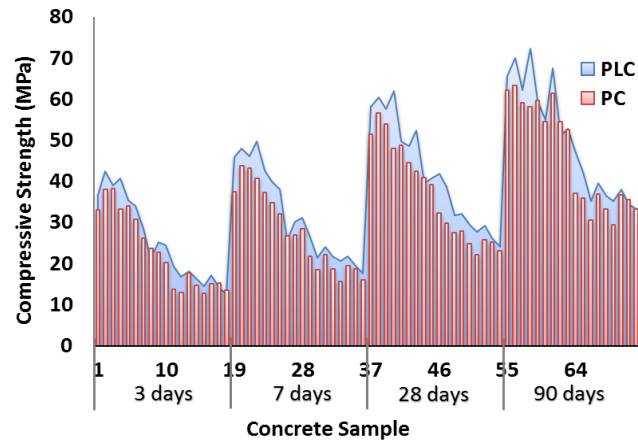
40

646
647
648   **Fig. 2.** The example ANN model (8-3-1)

649

650   **Fig. 3.** Experimental results for compressive strength of PC and PLC concrete

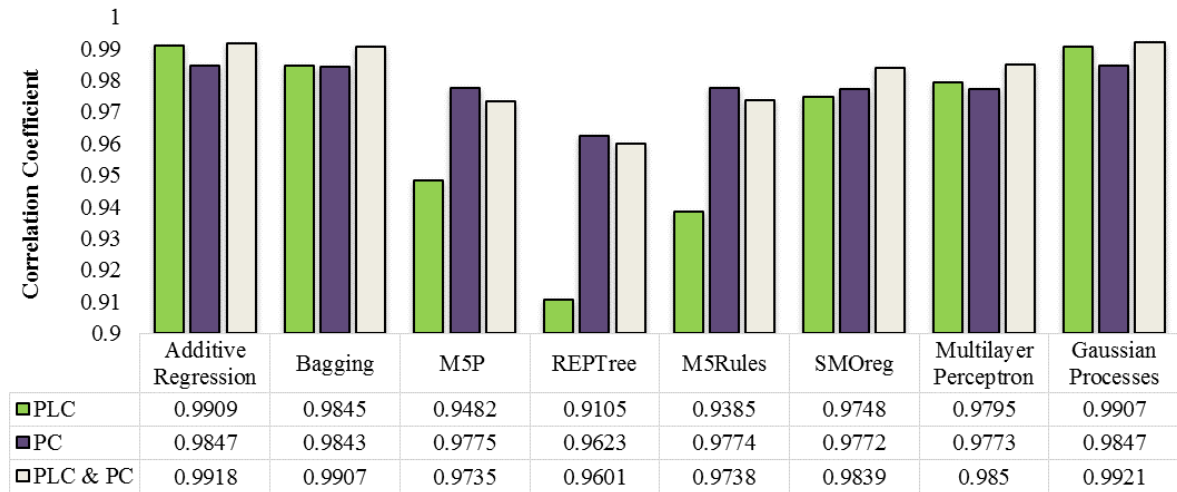|	Additive Regression	Bagging	M5P	REPTree	M5Rules	SMOreg	Multilayer Perceptron	Gaussian Processes
■ PLC	0.9909	0.9845	0.9482	0.9105	0.9385	0.9748	0.9795	0.9907
■ PC	0.9847	0.9843	0.9775	0.9623	0.9774	0.9772	0.9773	0.9847
□ PLC & PC	0.9918	0.9907	0.9735	0.9601	0.9738	0.9839	0.985	0.9921

651

652 **Fig. 4.** The highest R value for each of the models based on different datasets
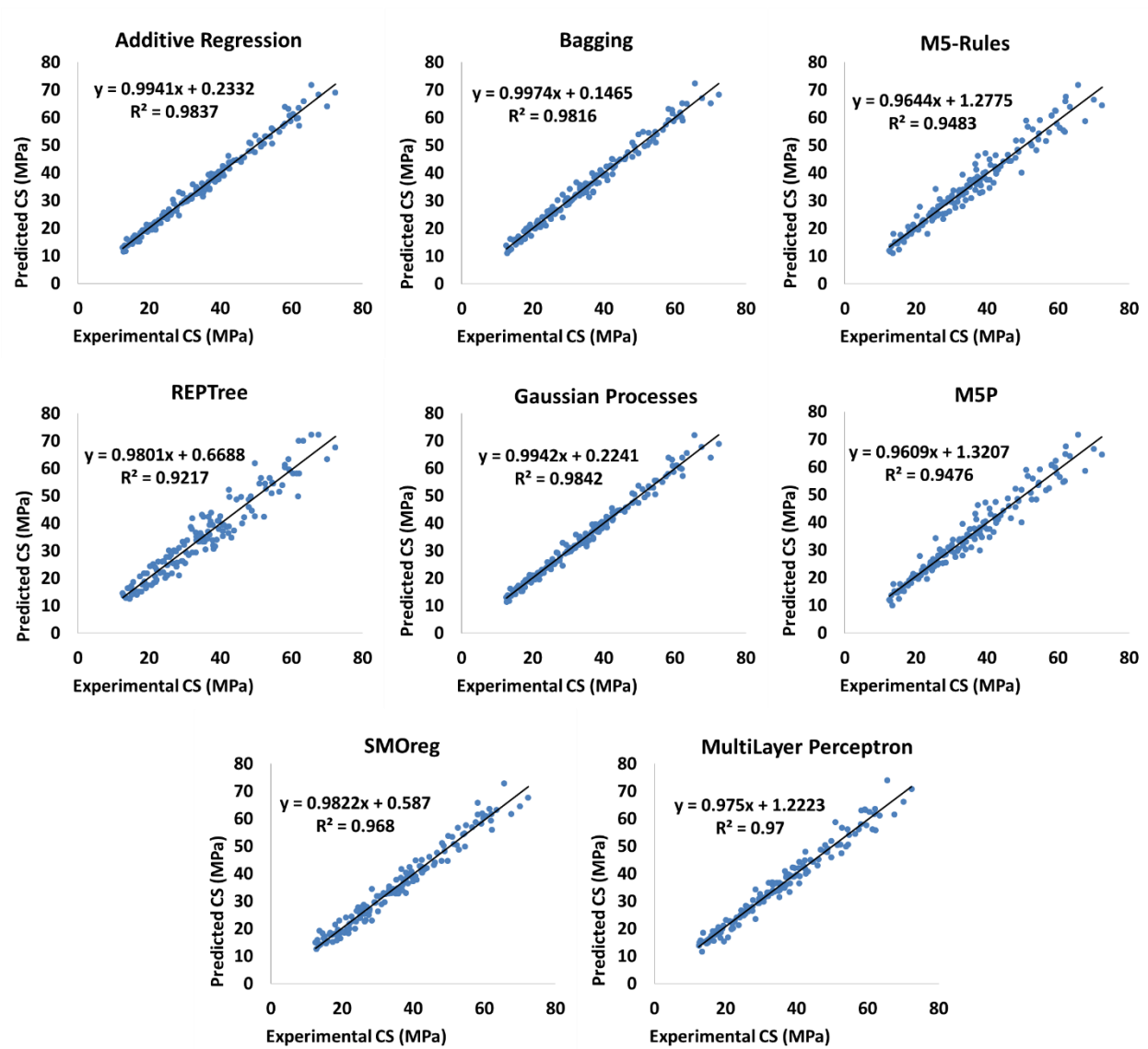
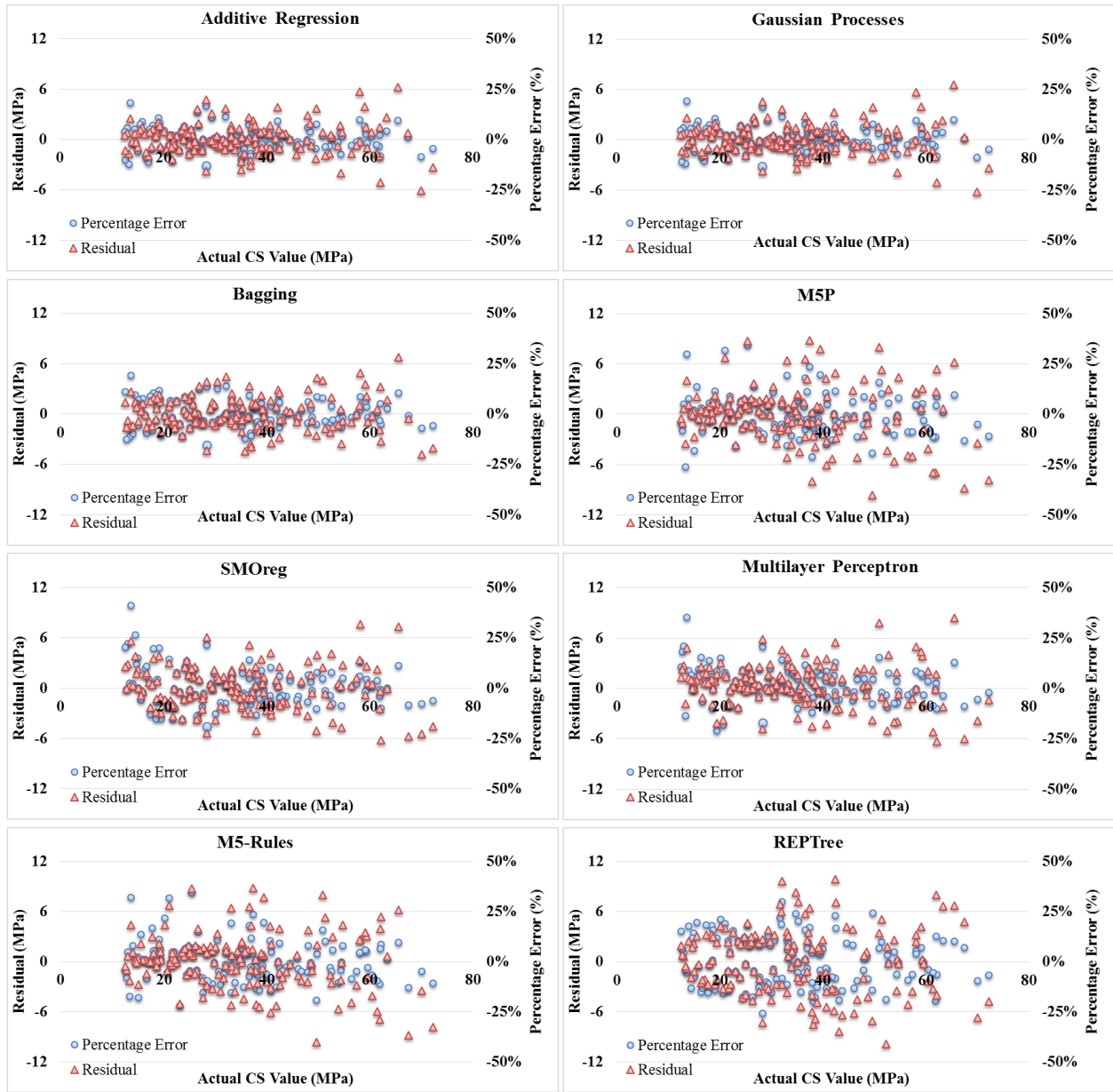**Fig. 5.** Predicted vs. actual compressive strength (abbreviated as CS in the figure)

653

654

Fig. 6. Residuals and percentage errors vs. actual compressive strength values

Sizing Worksheet (.xls)
Click here to download Sizing Worksheet (.xls): Sizing Worksheet.xlsx

ASCE Worksheet for Sizing Technical Papers & Notes

Please complete and save this form then email it with each manuscript submission.

Note: The worksheet is designed to automatically calculate the total number of printed pages when published in ASCE two-column format.

Journal Name:	**J. Computing in Civil Engineering**	Manuscript # (if known):	
Author Full Name:	B. Abounia-Omran, Q. Chen, & R. Jin	**Author Email:**	chen.1399@osu.edu

The maximum length of a technical paper is 10,000 words and word-equivalents or 8 printed pages. A technical note should not exceed 3,500 words and word-equivalents in length or 4 printed pages. Approximate the length by using the form below to calculate the total number of words in the text and adding it to the total number of word-equivalents of the figures and tables to obtain a grand total of words for the paper/note to fit ASCE format. Overlength papers must be approved by the editor; however, valuable overlength contributions are not intended to be discouraged by this procedure.

1. Estimating Length of Text

A. Fill in the four numbers (highlighted in green) in the column to the right to obtain the total length of text.

NOTE: Equations take up a lot of space. Most computer programs don't count the amount of space around display equations. Plan on counting 3 lines of text for every simple equation (single line) and 5 lines for every complicated equation (numerator and denominator).

2. Estimating Length of Tables

A. First count the longest line in each column across adding two characters between each column and one character between each word to obtain total characters.

1-column table = up to 60 characters wide	2-column table = 61 to 120 characters wide

B. Then count the number of text lines (include footnote & titles)

1-column table = up to 60 characters wide by:	**2-column table = 61 to 120 characters wide by:**
17 lines (or less) = 158 word equiv.	17 lines (or less) = 315 word equiv.
up to 34 lines = 315 word equiv.	up to 34 lines = 630 word equiv.
up to 51 lines = 473 word equiv.	up to 51 lines = 945 word equiv.
up to 68 text lines = 630 word equiv.	up to 68 text lines = 1260 word equiv.

C. Total Characters wide by Total Text lines = word equiv. as shown in the table above. **Add word equivalents** for each table in the column labeled "**Word Equivalents.**"

3. Estimating Length of Figures

A. First reduce the figures to final size for publication.

Figure type size can't be smaller than 6 point (2mm).

B. Use ruler and measure figure to fit 1 or 2 column wide format.

1-column fig. = up to 3.5 in.(88.9mm)	2-col. fig. = 3.5 to 7 in.(88.9 to 177.8 mm) wide

C. Then use a ruler to check the height of each figure (including title & caption).

1-column fig. = up to 3.5 in.(88.9mm) wide by:	**2-column fig. = 3.5 to 7 in.(88.9 to 177.8 mm) wide by:**
up to 2.5 in.(63.5mm) high = 158 word equiv.	up to 2.5 in.(63.5mm) high = 315 word equiv.
up to 5 in.(127mm) high = 315 word equiv.	up to 5 in.(127mm) high = 630 word equiv.
up to 7 in.(177.8mm) high = 473 word equiv.	up to 7 in.(177.8mm) high = 945 word equiv.
up to 9 in.(228.6mm) high = 630 word equiv.	up to 9 in.(228.6mm) high = 1260 word equiv.

D. Total Characters wide by Total Text lines = word equiv. as shown in the table above. **Add word equivalents** for each table in the column labeled **"Word Equivalents."**

Total Tables/Figures:	4887	(word equivalents)
Total Words of Text:	8289	

Total words and word equivalents:	**13176**
printed pages:	**11**

Estimating Length of Text

Count # of words in 3 lines of text:	38
Divided by 3	3
Average # of words per line	13
Count # of text lines per page	23
# of words per page	291.33
Count # of pages (don't add references & abstract)	19.5
Title & Abstract	500
Total # refs 63	1527
Length of Text is	7708 subtotal
	581 plus headings
	8289 TOTAL words
	7 printed pages

Estimating Length of Tables & Figures:

Tables	Word Equivalents	Figures	Word Equivalents
Table 1	158	Figure 1	158
2	158	2	158
3	158	3	158
4	158	4	315
5	630	5	945
6	158	6	945
7	158	7	
8	630	8	
9		9	
10	0	10	
11	0	11	
12	0	12	
13	0	13	
14	0	14	
15	0	15	
		16	
Please double-up tables/figures if additional space is needed (ex. 20+21).		17	
		18	
		19	
		20 and 21	

updated 1/16/03