A weighted multi-output neural network model for the prediction of rigid pavement deterioration

Fengdi Guo [a]*, Xingang Zhao [b], Jeremy Gregory [a] and Randolph Kirchain [c]

[a] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA;* [b] *Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA;* [c] *Materials Research Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA*

*Corresponding author: Fengdi Guo
Email: guofd@mit.edu
Department of Civil & Environmental Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue, Building E19-695
Cambridge, MA 02139, USA

A weighted multi-output neural network model for the prediction of rigid pavement deterioration

Abstract: A novel weighted multi-output neural network (NN) model is proposed for predicting the deterioration of rigid pavements based on Iowa pavement management system data. This first-of-a-kind model simultaneously predicts four pavement condition metrics concerning rigid pavements, including IRI, faulting, longitudinal crack and transverse crack. It provides an opportunity to efficiently evaluate pavement conditions and to make treatment decisions based on multi-condition metrics, such as the pavement condition index (PCI) for budget allocation models. Compared to traditional single-output NN models, this multi-output model is capable of incorporating correlations among different condition metrics. During model training, each condition metric is assigned a weight to reflect its relative importance. When the weights equal to those in the formula for the multi-condition metric, the prediction performance for PCI is optimal (13% lower MSE than optimal, single-output models). The multi-output model improves on the prediction performance for three of the four individual condition metrics compared to optimal single-output models. Results show that the consideration of correlations could improve the prediction performance for single and multi-condition metrics. Finally, variable weighting is critical for achieving the optimal balance of prediction performance among the various metrics as dictated by the needs of the decisionmaker.

Keywords: multi-output; neural network; deterioration; rigid pavement; correlation

1. Introduction

Performance-based planning (PBP) is an important tool to mitigate the pervasive problem of inadequate budget faced by transportation agencies (Baker *et al.* 2006). A key element for implementing PBP is efficient prediction of future pavement conditions. This depends on a robust deterioration prediction model. Over the last decades, many deterioration models have been described in the literature. Given the complex physical and chemical processes involved in pavement deterioration, most existing deterioration models are data-driven using many statistical methods including Markov chains (Kobayashi *et al.* 2012, Lethanh and Adey 2013), linear/nonlinear regression (Prozzi and Madanat 2004, Mahmood *et al.* 2018), neural network (Choi *et al.* 2004, Hossain *et al.* 2019), and decision trees (Inkoom *et al.* 2019).

In practice, PBP-based decisions are based on multiple metrics of pavement condition. These can include metrics of pavement roughness (e.g., the international roughness index (IRI)), rutting, faulting, and various forms of cracking (American Association of State Highway and Transportation Officials 2008, Bektas, Fatih; Smadi, Omar G.; and Al-Zoubi 2014, FHWA 2019). To date, deterioration models described in the literature predict single metrics (there are papers that describe multiple models for multiple metrics, but each metric is projected using a distinct single model). As will be shown in this paper, by examining panel data from the Iowa Pavement Management System (PMS) (which will be discussed in Section 3), the authors have observed that several of these metrics evolve in a coordinated manner. Failure to account for this correlation may lead to misestimation of future pavement condition, particularly in a stochastic context.

In making PBP-based decisions, transportation agencies may also use an overall condition metric – that is a weighted composite metric – to evaluate pavement performance. For example, the state of Iowa uses the pavement condition index (PCI), which is a weighted function of IRI, faulting, longitudinal and transverse cracking for rigid pavements (Bektas, Fatih; Smadi, Omar G.; and Al-Zoubi 2014). As will be shown in this paper, the nature of this weighting should be considered in the development of a correlation-aware prediction model.

In this paper, a new weighted multi-output neural network model for rigid pavements is proposed and evaluated using empirical data from the Iowa Pavement Management System. Considering the importance of data cleaning for the training of a data-driven model, a detailed description for the data cleaning process is presented. This model simultaneously predicts four individual metrics of pavement condition and a composite PCI. Different from multiple single-output models, this multi-output model can and does incorporate correlations among different condition metrics. Results show that after incorporating correlations among different condition metrics based on the weighed multi-output model, the prediction performance of the overall condition metric has been improved. In addition, the prediction performances of individual condition metrics are equal to or better than the performance of single-output models. Finally, by adding weights for each condition metric during the model training process, their importance levels can be tailored to the specific decision-making context.

This paper is organized as follows. Firstly, a comprehensive literature review is provided. Secondly, in the Data Preparation section, the processes of data collection and cleaning are described in detail. Next, the weighted multi-output neural network is introduced, including the selection of model input parameters, the normalization of training data, the model framework for rigid pavements and the model training process. Then the results about the comparison among single-output models and multi-output models with different weights are discussed. Finally, the main conclusions are summarized, and recommended future research activities are presented.

2. Literature Review

For data-driven models, the first methodological consideration is selection of the model form and the method to estimate model parameters in the deterioration model. As suggested in (Swei *et al.* 2018), deterioration modeling methods can be divided into two categories, Markov chains and regression models. Models based on Markov chains focus on the probability that pavement condition evolves from one state to another state (Hong and Wang 2003, Kobayashi *et al.* 2012, Lethanh and Adey 2013, Abaza 2016). This type of models usually assumes the variation of pavement deterioration to be aleatory. Since the Markov property assumes that future deterioration rates only depend on the current pavement condition (i.e. pavement state), historical dependence is ignored during the prediction process.

There are several types of regression models, among which the most commonly applied in the deterioration literature is linear regression (including both explicitly linear

forms (Mahmood *et al.* 2018) and linear forms by transformation[1] (Prozzi and Madanat 2003, 2004, Swei *et al.* 2018)). These models can be readily solved by ordinary least squares (OLS). In order to consider the heterogeneity of individual pavements in the same pavement group, Zhang et al. proposed a clusterwise linear regression model (Zhang and Durango-Cohen 2013), while Yu et al. proposed linear mixed effects models (Yu *et al.* 2007). When the prediction is concerned with classification or the probability of performance failure occurrence, logistic regression can be applied. Heidari et al. applied a logistic regression to predict the subclasses of pothole, rutting and protrusion for forest roads (Heidari *et al.* 2018). Chen et al. evaluated the failure probability of asphalt preventive treatments by both logistic and Bayesian logistic models (Chen *et al.* 2019).

The second type of regression models are non-linear models which are described by the sum of several nonlinear variables or by the exponential form. AASHTO models are examples of this type that were once widely applied and cited. These models were estimated from the AASHO road test data and updated for several versions (Highway Research Board 1962, American Association of State Highway and Transportation Officials 1993). Based on the first version of the AASHTO model, Ben-Akiva et al. developed a latent model for pavement deterioration, which assumed that the pavement performance condition is a latent variable (Ben-Akiva and Ramaswamy 1993). Abaza et al. proposed a deterministic prediction model based on the AASHTO serviceability concept and an incremental solution of the AASHTO basic design equation (Abaza 2004). Following the AASHTO model form, Hong et al. developed a nonlinear model that incorporated heterogeneity using Bayesian analysis (Hong and Prozzi 2006, 2010). The Mechanistic-Empirical Pavement Design Guide (MEPDG) is a collection of models that were developed as an update to AASHTO through the National Cooperative Highway Research Program (NCHRP) Project 1-37A (American Association of State Highway and Transportation Officials 2008). Considering the fact that the input parameters in the MEPDG model could be estimated by separate models and thus cause bias, Aguiar-Moya et al. introduced instrumental variables to update the MEPDG model (Aguiar-Moya *et al.* 2011). Other nonlinear models include the exponential form (Lamptey *et al.* 2008, Rosa *et al.* 2017) and multivariate adaptive regression splines (MARS) proposed by (Attoh-Okine *et al.* 2003).

Another class of regression models that have been applied to predict pavement performance are neural networks. A neural network model is an example of machine learning where a prediction algorithm is developed through iterations among a training set of values. The ultimate prediction algorithm combines both linear and non-linear elements weighted to produce a best prediction. Recent years have witnessed rapid expansion in the application of neural network models due to their efficiency to describe nonlinear relationships in many phenomena including pavements' deterioration (Nii O. Attoh-Okine 1994, Owusu-Ababio 1998, Roberts and Attoh-Okine 1998, Lou *et al.* 2001, Lee and Lee 2004, Thube 2012, Mazari and Rodriguez 2016, Gong *et al.* 2019, Inkoom *et al.* 2019). Neural network models usually consist of input layer, hidden layer and output layer. With the increase in model complexity, such as the increase of hidden layers and the increase of the neurons in each hidden layer, the neural network could fit the training data perfectly but may lack the

[1] Models like $f(\mathbf{x}) = \prod_i a_i x_i^{b_i}$ can be transformed into explicitly linear models by log-transformation: $\log f(\mathbf{x}) = \sum_i (\log a_i + b_i \cdot \log x_i)$

capacity to predict the future deterioration at the same time. This phenomenon is called 'overfitting'. There are several ways to detect and prevent it, such as out-of-sample prediction, cross-validation, regularization, and early-stopping, etc. However, some exiting models may lack the analysis to check if the neural network model is overfitting or not (Nii O. Attoh-Okine 1994, Owusu-Ababio 1998, Roberts and Attoh-Okine 1998, Mazari and Rodriguez 2016). These models focus more on the fitting process. Attoh-Okine (1994) and Roberts and Attoh-Okine (1998) mainly use the training error as the evaluation criteria for model performance, Owusu-Ababio (1998) and Mazari and Rodriguez (2016) just use 1-fold validation instead of multi folds (i.e. cross validation).

There are also some other types of regression models. Inkoom et al. proposed a classification and regression tree (CART) model to predict crack condition (Inkoom *et al.* 2019). Pan et al. used fuzzy regression to predict PSI by considering visual inspection data as fuzzy data (Pan *et al.* 2011). Bianchini et al. proposed a neuro-fuzzy model with the consideration of IF-THEN fuzzy rules (Bianchini and Bandini 2010).

Existing models mainly focus on the prediction of a single condition metric – a single overall metric like PCI, or a specific metric like IRI. This type of single-output models lacks the consideration of correlations among different condition metrics. They could also potentially increase the computational cost when generating multi specific metrics to obtain an overall metric, and lead to some inconvenience for a pavement management system whose treatment decisions are based on multi metrics.

The second methodological consideration is the selection of predictor (input) variables for deterioration models, which is strongly related with the available parameters in the PMS dataset available to the model developer. Common input variables include traffic volumes/loads, pavement structure (pavement type, thickness), pavement age, pavement conditions, and maintenance history. Environmental factors can also have a significant influence on the pavement deterioration, such as ambient temperature, precipitation, freeze-thaw cycles, and freeze index. Incorporating all parameters in regression models, on one hand could increase the model complexity and lead to overfitting, on the other hand it prevents us from understanding the role of each parameter. Several studies have conducted sensitivity analysis to explore the significance of input variables (Choi *et al.* 2004, Kargah-Ostadi *et al.* 2010, Karlaftis and Badr 2015, Hossain *et al.* 2019, Yao *et al.* 2019).

Models are inherently limited to the variables described in available data. Unfortunately, several current PMS datasets do not fully describe the maintenance history of a pavement segment, limiting the ability to model this effect. For example, many models use total pavement thickness as an input variable. However, two pavement segments with the same total thickness are very probable to have different deterioration rates: one asphalt segment has an original 12-inch (304.8 mm) thickness and does not have an overlay, the other one has an original 8-inch (203.2 mm) thickness and also has a 4-inch (101.6 mm) asphalt overlay.

Another important methodological consideration is the data "cleaning" process, which usually takes data scientists about 60% of the total time for a project (Press 2016). One common problem is the identification and processing of outliers in the dataset. For example, when there are no treatment actions, the pavement condition should get worse due to the traffic and environmental influence. However, in much real-world PMS data, like LTPP, overall pavement condition is sometimes recorded as getting better from one measurement to the next. There are several reasons for the existence of outliers, such as measurement error or

missing treatment records. Given the same raw dataset, with different cleaning criteria, different training datasets are obtained. When datasets differ, even the same training method can generate different models with different evaluation performance. This makes it challenging to compare different methods when cleaning methods are not reproducible. Most existing papers in the deterioration modeling literature have an obscure description about the data cleaning process.

To summarize, there are several gaps for existing pavement deterioration models described in the literature. The first gap is about the correlations among condition metrics. Current models are single-output ones that focus on either an overall or a specific condition metric. They lack the consideration of correlations among different condition metrics. The second gap is about input variable selection. Variables in most existing models may not reflect the maintenance history of a pavement segment. The last gap is the unclear description about the data cleaning process, leading to the difficulty of comparing different models.

To bridge current gaps in pavement deterioration models, an innovative weighted multi-output neural network model is proposed. Most existing models only focus on flexible pavements. According to the FHWA road statistics (FHWA 2017), rigid pavements account for over 25% in terms of interstate system. This paper uses rigid pavements as an example to illustrate this new multi-output neural network model. A detailed description of data cleaning process is presented. This new model incorporates correlations among different condition metrics during the model training process. Considering the fact that different metrics may have different importance levels for transportation agencies, each output condition metric is assigned a weight number instead of being uniformly weighted during the training process. Model evaluation results show that after incorporating correlations among different condition metrics, the prediction of the overall condition metric could be improved. In addition, the prediction performance of a single condition metric is better or equal to the performance of single-output models.

3. Data Preparation

The proposed deterioration model is trained on data from the Iowa PMS database and climate data from LTPP. In the following section, a detailed description is presented for the data preparation, which is concerned with data extraction, transformation and cleaning in terms of input and output variables. At the end of this section, the selection of input variables is discussed based on a correlation analysis.

3.1 Iowa PMS data

The Iowa PMS database contains detailed records for around 4,000 pavements segments on 26 years of pavement condition for the years 1992 to 2017. It covers three systems: 1,566 miles (2,520 km) for interstate system, 4,626 miles (7,445 km) for U.S route system, and 4,891 miles (7,871 km) for Iowa route system. Approximately, 15.6% of total miles are asphalt pavements, 32.9% are concrete pavements, and 51.5% are composite pavements.

This data can be divided into three groups based on the units used and the presence of crack information as shown in Table 1. Because the early period of data (Group 1) contains different information about the state of cracking, it is not considered in this analysis. In the dataset, there are two variables called ‘IRIDAT’ and ‘CAPDAT’ which record the dates of measurement for the IRI and crack information, respectively. For each segment in the

network, the performance data is collected biennially. For the dataset of year 2013, 'CAPDAT' is missing, and for the dataset of year 2016 and 2017, 'IRIDAT' is missing. Therefore, the PMS data that is used for training the deterioration model is based on year 2000-2012, year 2014 and 2015.

Table 1. Iowa PMS data groups

Group	Period	Units	Crack info
1	1992-1999	1992-1995: English, 1996-1999: Metric	DCRACK, HCRACK, LCRACK, TCRACK
2	2000-2012	Metric	ACRACK, LCRACK, LWCRACK, TCRACK
3	2013-2017	English	ACRACK, LCRACK, LWCRACK, TCRACK

* ACRACK: alligator crack, DCRACK: durability crack, HCRACK: half-crack, LCRACK: longitudinal crack, LWCRACK: longitudinal wheel-path crack, TCRACK: transverse crack

The detailed information for parameters in the selected PMS data can be found in (Iowa DOT Open Data 2019). Based on suggestions from engineering experts, four types of parameters have been selected in this analysis as listed in Table 2. Parameters in the *Basic Info* group include the segment name (ORIGKEY), the year for the database (PMISYR), system type (SYSTEM, including interstate, US route, and Iowa route), the construction year (CONSYR) and the year for last resurfacing (RESYR). Group *Traffic* includes two types of traffic statistics, AADT and AADTT. Different kinds of condition metrics and corresponding measurement dates are in Group *Distress*. The last group *Maintenance History* records treatment information for each layer.

Table 2. Description of selected parameters.

Data type	Name	Description
Basic Info	ORIGKEY	Original smart key
	PMISYR	Pavement management year
	SYSTEM	Roadway system
	PAVTYP	Pavement type
	CONYR	Year of construction or reconstruction
	RESYR	Year of last resurfacing
Traffic	AADT	Average annual daily traffic
	TRUCKS	Average annual daily truck traffic
Distress	IRI	International roughness index
	RUT	Rut depth
	FAULT	Average faulting only on faulted joints in a segment
	ACARCKH, ACRACKM, ACRACKL	High/moderate/low severity alligator cracks
	LCARCKH, LCRACKM, LCRACKL	High/moderate/low severity longitudinal cracks
	LWCARCKH, LWCRACKM, LWCRACKL	High/moderate/low severity longitudinal wheel-path cracks
	TCARCKH, TCRACKM, TCRACKL	High/moderate/low severity transverse cracks
	IRIDAT	IRI test year
	CAPDAT	Crack & patch collected by vender test year

	TREATMENT	Surface treatment type
	LAYR (1-8)	Layer year # (1-8)
	SURTYP (1-8)	Surface type # (1-8)
	SURTHK (1-8)	Surface thickness # (1-8)
Maintenance	BASTYP (1-8)	Base type # (1-8)
History	BASTHK (1-8)	Base thickness # (1-8)
	SUBTYP (1-8)	Subbase type # (1-8)
	SUBTHK (1-8)	Subbase thickness # (1-8)
	RMVTYP (1-8)	Removal type # (1-8)
	RMVTHK (1-8)	Removal thickness # (1-8)

As for Basic Info parameters, based on 'PMISYR', 'CONYR' and 'RESYR', two new variables 'AGECON' and 'AGERES' were generated by equation (1) and (2). These two variables were used to describe the period since construction/reconstruction and last resurfacing, respectively.

$$\mathrm{AGECON} = \mathrm{PMIYR} - \mathrm{CONYR} \tag{1}$$

$$\mathrm{AGERES} = \mathrm{PMIYR} - \mathrm{RESYR} \tag{2}$$

Distress parameters in year 2014 and 2015 were transformed into metric units. Each type of crack distress is described by three sub-categories, which can be converted to a single crack distress metric by equation (3) (Bektas, Fatih; Smadi, Omar G.; and Al-Zoubi 2014). Using this, four new variables can be obtained, namely: 'ACRACK', 'LCRACK', 'LWCRACK' and 'TCRACK'.

$$\mathrm{CRACK} = \mathrm{CRACKL} + 1.5\mathrm{CRACKM} + 2\mathrm{CRACKH} \tag{3}$$

According to the Iowa Department of Transportation, overall pavement condition is not evaluated directly from condition metric (e.g., IRI). Instead, condition metrics are transformed into indexes including the cracking index, riding index (IRI), and faulting index (Bektas, Fatih; Smadi, Omar G.; and Al-Zoubi 2014). Table 3 lists the threshold values for the calculation of indexes for rigid pavements. Each metric has two threshold values. When the metric ζ is less than or equal to the small threshold value ζ_*, the index equals 100; When the metric is larger than or equal to the large threshold value ζ^*, then the index equals 0. Other index values can be obtained by linear interpolation. Following this, for any metric ζ, the corresponding index, I^{ζ}, is defined formally as:

$$I^{\zeta} = \begin{cases} 100 & \zeta \le \zeta_* \\ \dfrac{\zeta - \zeta^*}{\zeta_* - \zeta^*} \cdot 100 & \zeta_* < \zeta < \zeta^* \\ 0 & \zeta^* \le \zeta \end{cases} \tag{4}$$

An integrated cracking index was obtained as a weighted sum of indexes of LCRACK and TCRACK with the corresponding weights listed in Table 4.

Table 3. Thresholds for condition metrics

	IRI (m/km)	**FAULT (mm)**	**LCRACK (m/km)**	**TCRACK (count/km)**
Threshold	0.5&4	0&12	0&250	0&150

Table 4. Cracking sub-index weights

sub-index	LCRACK	TCRACK
weights	0.4	0.6

Based on these indexes, an overall condition metric called PCI was obtained by equation (5) for rigid pavements. PCI ranges from 0 to 100, and 100 represents that pavement is in perfect condition. In this expression, each of the condition indexes have different coefficients which may represent the significance of that index in the eyes of the Iowa transportation agency.

$$PCI = 0.4I^{Crack} + 0.4I^{Ride} + 0.2I^{Fault} \quad (5)$$

The last group of parameters describe the maintenance history for each segment, from which historical structure information could be generated, including construction year, resurface year, thickness and pavement type. Table 5 presents an example for the maintenance history for a rigid pavement segment. Its ORIGKEY is "08012183 67192 8279". This segment was constructed in 1964 and maintained with an overlay of Portland cement concrete (PCC) in 1984. Based on Table 5, structure information could be obtained as shown in Table 6. From year 1964 to year 1983, the segment structure stayed the same, AGECON (equation (1)) changes each year and AGERES remained 0. After year 1984, due to the overlay action, the segment structure was changed. The new surface thickness SURTHK is 102mm, and previous surface PCC thickness become PCCTHK. TOTTHK is the total thickness, which is equal to the sum of SURTHK and PCCTHK (RMVTHK should be considered when it exists). MRRNUM is the number of treatment actions. In this case, MRRNUM became 1 since 1984. Instead of a single total thickness value as shown in the original PMS dataset, the introduction of SURTHK, PCCTHCK and MRRNUM could reflect the maintenance history of a pavement segment. (If the pavement type is asphalt or composite pavements, a variable called 'HMATHK' is introduced similar to PCCTHK.)

Table 5. Maintenance history for a rigid pavement segment

LAYR	SURTYP	SURTHK	BASTYP	BASTHK
1964	PCC	254	GSB	102
1984	PCC	102	-	-

*GSB: granular subbase

Table 6. Structure information for a rigid pavement segment

PMISYR	CONYR	RESYR	PAVTYP	TOTTHK	SURTHK	PCCTHK	MRRNUM
1964~1983	1964	-	PCC	254	254	0	0
>=1984	1964	1984	PCC	356	102	254	1

After generating parameters for model training, the total dataset is filtered based on 'PAVTYP' to obtain the data for rigid pavements. Next, input and output variables are determined with the consideration of time series analysis and treatment actions. In terms of the time series analysis, time lag = 1 is suggested in (Chu and Durango-Cohen 2008, Swei *et al.* 2018), which means that the condition prediction at year t+1 is only related with the information at year t. Since the condition metrics are collected every two years in Iowa, time lag is selected as 2 years here. For example, the information at year 2000 can be considered as inputs and the condition metrics at year 2002 are outputs.

During the process of paring input and output variables, attention should be paid to pavement treatments. If there is a treatment action at year 2000, then it is not appropriate to use year 2000 as input and 2002 as output. This kind of pairing data should be excluded from the training dataset. In the dataset, the improvement of pavement condition may be postponed after a treatment. For example, there is a treatment at year 2002, but the condition improvement is reflected at year 2004. Therefore, during the pairing process, when there exists a treatment action, its adjacent years are checked to decide which year should be considered as the treatment year.

After this step, input variables include: 1). Basic info: AGECON, AGERES; 2). Traffic: ADT and TRUCKS; 3). Distress: IRI_t, FAULT_t, LCRACK_t, TCRACK_t; 4). Maintenance history: TOTTHK, SURTHK, PCCTHK, MRRNUM. Output variables include IRI_{t+2}, FAULT_{t+2}, LCRACK_{t+2}, TCRACK_{t+2}. The subscription *t* and *t*+2 represent measurement years of condition metrics.

3.2 Iowa PMS data cleaning

After generating the initial input and output variables, the next step is to deal with data outliers. For this case, outliers are defined as values or pairs of values that indicate physically unreasonable pavement condition values. In terms of the overall pavement condition (i.e. PCI in this analysis), it should decrease when there is no treatment action due to running traffic and environmental effects (American Association of State Highway and Transportation Officials 2008). But for other single performance metrics (IRI, faulting and cracks), their performances may improve due to some negative correlations. Differences for all performance metrics between input and output are calculated,

$$\Delta\mathrm{PCI} = \mathrm{PCI}_{t+2} - \mathrm{PCI}_t \tag{6}$$

$$\mathrm{IRI\_diff} = \mathrm{IRI}_{t+2} - \mathrm{IRI}_t \tag{7}$$

$$\mathrm{FAULT\_diff} = \mathrm{FAULT}_{t+2} - \mathrm{FAULT}_t \tag{8}$$

$$\mathrm{LCRACK\_diff} = \mathrm{LCRACK}_{t+2} - \mathrm{LCRACK}_t \tag{9}$$

$$\mathrm{TCRACK\_diff} = \mathrm{TCRACK}_{t+2} - \mathrm{TCRACK}_t \tag{10}$$

When $\Delta\mathrm{PCI}$ is larger than 0, the data point is considered as an outlier and is deleted from the training dataset. In terms of other single performance metrics, interquartile range (IRQ) is applied, which can be calculated by equation (11),

$$IQR = Q_3 - Q_1 \tag{11}$$

Where Q_1 and Q_3 represent values of 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, respectively. Usually, if a data point is below $Q_1 - 1.5 \cdot IQR$ or above $Q_3 + 1.5 \cdot IQR$, it is considered as an outlier (Rice 2006). In this case, all lower quartile boundaries are negative values, indicating unexplained performance improvement. These lower bounds have been used to identify outliers. There is less theoretical support to suggest that pavements could not deteriorate rapidly. Therefore, an upper bound of one half of the threshold values has been adopted for use in Table 3, namely 2 for IRI. Table 7 shows the outlier threshold for condition metrics based on *IQR* and the half

of index approach. Values outside the lower and upper bound were eliminated from the training dataset.

Table 7. Outlier threshold for condition metrics

	IRI_diff	Fault_diff	LCRACK_diff	TCRACK_diff
$Q_1 - 1.5 \cdot IQR$	-0.3	-4.9	-23.0	-13.5
$Q_3 + 1.5 \cdot IQR$	0.4	6.0	34.4	18.5
Half of index threshold	2.0	6.0	125.0	75.0

3.3 Climate data

Climate data was primarily extracted from LTPP, including annual average temperature (TEMP), precipitation (PRECIP), and freeze index (FREEZE). For a given year, there are several records for each climate variable. Annual averages of these were used. Climate data was matched to the other training data (which are generated in section 3.1 and 3.2) by condition metric measurement year, IRIDAT and CAPDAT.

4. Weighted multi-output neural network model

4.1 Selection of input parameters

Based on the previous analysis, a candidate set of input variables were generated. However, some variables may not contribute to the pavement deterioration process to an extent that is observable in the current data. To select the most relevant input parameters, a correlation analysis was conducted. This approach is similar to the linear regression analysis described by (Kargah-Ostadi *et al.* 2010).

Fig 1 shows the heatmap for the correlations among input and output variables. The output variables are the variations of condition metrics, including IRI_diff, FAULT_diff, LCRACK_diff, and TCRACK_diff (equations (7)-(10)). All other listed variables are candidate inputs. In this analysis, an input variable would be excluded if either the absolute value of the correlation coefficient is less than 0.05 for all four outputs or if it is less than 0.05 for three outputs and less than 0.1 for the fourth. Using these criteria for this dataset, AGERES, PCCTHK, MRRNUM and TEMP were excluded. One main reason that output response is not correlated with AGERES, PCCTHK, MRRNUM is that for most rigid pavements in Iowa the historical overlay number is zero across the dataset. When overlay number is zero, AGERES, PCCTHK, MRRNUM are all equal to zero, and TOTTHK is equal to TOPTHK for most rigid pavements. To simply the problem, AGERES, PCCTHK, and MRRNUM are deleted from the set of input variables. But to be noted that, for asphalt and composite pavements, MRRNUM is usually larger than 0, these variables should be considered. ADT is insignificant for both variations of IRI (IRI_diff) and longitudinal crack (LCRACK_diff). In addition, it is also strongly correlated with TRUCKS. Hence, ADT is also deleted.

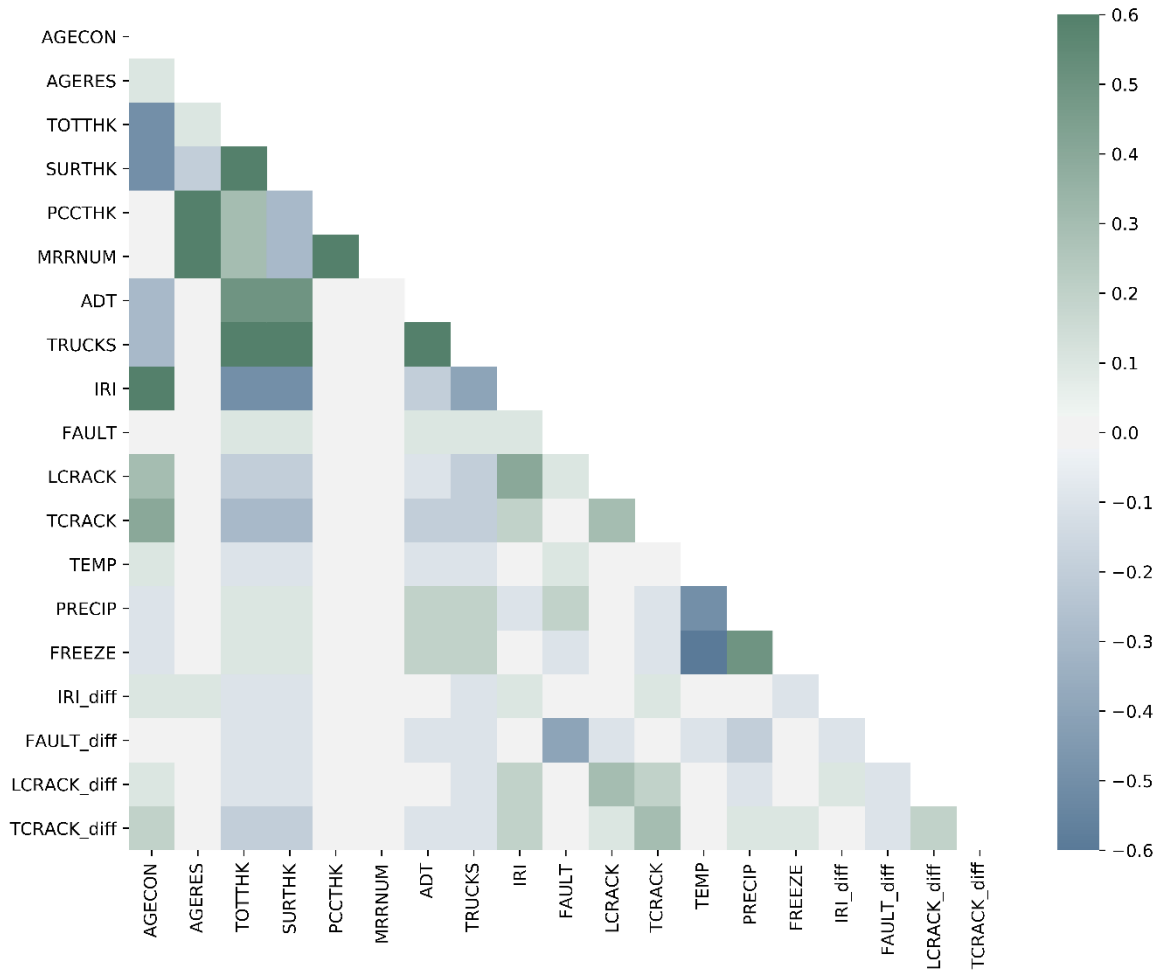


Figure 1. Heatmap for variables' correlation, green color represents positive correlation, blue represents negative correlation, and grey color represents weak correlation

Among the four outputs, several interesting correlation trends can be observed. First, the variation of IRI (IRI_diff) is positively correlated with both the level of IRI (collection coefficient, $\rho = 0.13$) and the level of transverse cracking (TCRACK, $\rho = 0.08$). For this modeling, a positive correlation suggests that IRI increases more rapidly when values for IRI or TCRACK are higher. Additionally, it has been noticed that the variation of FAULT (FAULT_diff) is strongly negatively corelated with the level of faulting (FAULT, $\rho = -0.36$) and longitudinal cracking (LCRACK, $\rho = -0.12$). This suggests that faulting grows more slowly when faulting or longitudinal cracking is more severe. The variation of LCRACK (LCRACK_diff) and TCRACK (TCRACK_diff) are both strongly positively correlated with IRI, LCRACK and TCRACK. From this correlation analysis, it can be concluded that the deterioration of each single condition metric is statistically dependent on at least some of the other condition metrics. This observation implies that it's better to incorporate the state of all condition metrics during the prediction of single condition metrics.

Table 8 summarizes the distribution of input variables for the neural network model. As for four condition metrics, when their standard deviation and mean values are compared, IRI has the smallest ratio between standard deviation and mean value, representing the IRI distribution is more centralized. TCRACK has the largest ratio and its distribution is the sparsest, which might be caused by the measurement quality.

Table 8. Statistical summary for input variables

	AGECON	TOTTHK	TRUCKS	IRI	FAULT	LCRACK	TCRACK	PRECIP	FREEZE
mean	22.2	248.8	1660	2.2	4.6	19.6	13.4	820.8	286.8
std	14.9	32.5	2559.7	0.8	2.6	37.8	26.8	112.2	156.6
min	0	150	1	0.7	0	0	0	650.5	317.8
max	71	559	14293	5	12	350	240.5	1077.2	840.8

*std: standard deviation

4.2 Normalization of training data

As listed in Table 8, each variable has a different magnitude. To avoid fitting problems due to this, all input and output variables are normalized. Considering that both input and output variables have condition metrics but for different years, all condition metrics are normalized together. Suppose **X** is the input matrix with the dimension is $d_x \times n$, d_x is the number of input variables, and n is the number of data points. **Y** is the output matrix. Its dimension is $d_y \times n$, and d_y is the number of output variables. The normalization process is shown as follows:

$$x_{j,i} = \frac{X_{j,i} - \overline{X}_{j,}}{\sigma_{X_{j,}}} \quad \left(j \notin \left\{\text{IRI, FAULT, LCRACK, TCRACK}\right\}\right) \tag{12}$$

$$x_{j,i} = \frac{X_{j,i} - \overline{Y}_{j,}}{\sigma_{Y_{j,}}} \quad \left(j \in \left\{\text{IRI, FAULT, LCRACK, TCRACK}\right\}\right) \tag{13}$$

$$y_{k,i} = \frac{Y_{k,i} - \overline{Y}_{k,}}{\sigma_{Y_{k,}}} \quad \left(k \in \left\{\text{IRI, FAULT, LCRACK, TCRACK}\right\}\right) \tag{14}$$

Where $x_{j,i}$ and $y_{k,i}$ are normalized input and output values, where *i* represents the index of data point, *j* is the input variable index and *k* is the output variable index. $\overline{X}_{j,}$ and $\overline{Y}_{k,}$ are mean values for j_{th} input and k_{th} output. $\sigma_{X_{j,}}$ and $\sigma_{Y_{k,}}$ are standard deviations for j_{th} input and k_{th} output.

4.3 Model framework

In recent years, feed-forward neural network (NN) models have been widely applied for pavement deterioration prediction due to their superior performance when dealing with strongly nonlinear relationships (Karlaftis and Badr 2015, Mazari and Rodriguez 2016, Heidari *et al.* 2018, Hossain *et al.* 2019, Inkoom *et al.* 2019, Yao *et al.* 2019). A NN model usually consists of an input layer, one or several hidden layers, and an output layer. Each layer has several neurons. Hidden layers are usually used to describe the nonlinear relationships via activation functions (also known as transfer functions). The connection between two layers are usually through a linear mapping. A NN model can be mathematically described as equation (15)

$$\hat{\mathbf{Y}} = \sigma_{H+1}\left(\mathbf{W}^{H+1}\sigma_{H}\left(\mathbf{W}^{H}\sigma_{H-1}\left(\mathbf{W}^{H-1}...\sigma_{2}\left(\mathbf{W}^{2}\sigma_{1}\left(\mathbf{W}^{1}\mathbf{X}\right)\right)\right)\right)\right) \tag{15}$$

where $\hat{\mathbf{Y}}$ represents the prediction values. $\mathbf{W}^{h}\left(h=1,\ldots H+1\right)$ is the linear transformation matrix, and $\sigma_{h}\left(h=1,\ldots H+1\right)$ is the activation function. Common nonlinear activation functions include ReLU, sigmoid, and Tanh. Their ouptut sets are $\left[0,\infty\right]$, $\left(0,1\right)$ and $\left(-1,1\right)$. ReLU is often used for regression problems while the other two are more commonly applied for classification problems (Bishop 2006). In terms of the current regression problem, ReLU (equation (16)) is applied for the hidden layers and a linear function is applied for the output layer.

$$f(x)=\begin{cases} x & x>0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

The goal of training a NN model is to determine an optimal structure that can minimize the loss funcion. A loss function is used to describe the relationship between the predicted value $\hat{\mathbf{Y}}$ and the true value $\mathbf{Y}$. In this regression analysis, mean squared error is chosen as the loss function (equation (17)), where λ_k^2 is the weight value for each output variable.

$$\min: L\left(\hat{\mathbf{Y}},\mathbf{Y}\right)=\frac{1}{2}\sum_{k=1}^{d_y}\lambda_k^2\left\|\hat{\mathbf{Y}}_k-\mathbf{Y}_k\right\|_2^2=\frac{1}{2}\sum_{k=1}^{d_y}\sum_{i=1}^{n}\lambda_k^2\left(\hat{Y}_{k,i}-Y_{k,i}\right)^2 \tag{17}$$

After normalization process introduced in section 4.2, the training problem for a NN model (equation (15) and (17)) can be reformulated as,

$$\hat{\mathbf{y}} = \sigma_{H+1}\left(\mathbf{W}^{H+1}\sigma_{H}\left(\mathbf{W}^{H}\sigma_{H-1}\left(\mathbf{W}^{H-1}...\sigma_{2}\left(\mathbf{W}^{2}\sigma_{1}\left(\mathbf{W}^{1}\mathbf{x}\right)\right)\right)\right)\right) \tag{18}$$

$$\min: L\left(\hat{\mathbf{y}},\mathbf{y}\right)=\frac{1}{2}\sum_{k=1}^{d_y}\lambda_k^2\left\|\hat{\mathbf{y}}_k-\mathbf{y}_k\right\|_2^2=\frac{1}{2}\sum_{k=1}^{d_y}\sum_{i=1}^{n}\lambda_k^2\left(\hat{y}_{k,i}-y_{k,i}\right)^2 \tag{19}$$

In this case, the prediction value is $\hat{\mathbf{y}}$, which is in the normalization space. Equation (20) is applied to obtain the prediction value with a real magnitude.

$$\hat{y}_{k,i}=\frac{\hat{Y}_{k,i}-\overline{Y}_{k,}}{\sigma_{Y_{k,}}}, \text{ and } \hat{Y}_{k,i}=\overline{Y}_{k,}+\hat{y}_{k,i}\cdot\sigma_{Y_{k,}} \tag{20}$$

When the dimension of the output variable d_y is equal to 1, the NN is a single-output model, which is very common in existing literatures. When d_y is larger than 1, and $\lambda_k=1$ for all output variables, then this is a uniformly weighted multi-output model. When d_y is larger than 1, and λ_k are different for all output variables, it is a weighted multi-output NN model, which is the focus of this analysis.

In order to incorporate weights for each output variable, insert equation (14) to equation (19), then,

$$\min: L(\hat{\mathbf{y}},\mathbf{y}) = \frac{1}{2}\sum_{k=1}^{d_y}\sum_{i=1}^{n}\lambda_k^2\left(\hat{y}_{k,i}-y_{k,i}\right)^2 = \frac{1}{2}\sum_{k=1}^{d_y}\sum_{i=1}^{n}\lambda_k^2\left(\frac{\hat{Y}_{k,i}-\overline{Y}_{k,}}{\sigma_{Y_{k,}}}-\frac{Y_{k,i}-\overline{Y}_{k,}}{\sigma_{Y_{k,}}}\right)^2$$
$$= \frac{1}{2}\sum_{k=1}^{d_y}\sum_{i=1}^{n}\left(\frac{\hat{Y}_{k,i}-\overline{Y}_{k,}}{\sigma_{Y_{k,}}/\lambda_k}-\frac{Y_{k,i}-\overline{Y}_{k,}}{\sigma_{Y_{k,}}/\lambda_k}\right)^2 \quad (21)$$

By changing the stadard deviaiton during the normalization process, weights in the loss function can be incorporated.

4.4 Weighted multi-output NN model for rigid pavements

In terms of the NN model structure for rigid pavement deterioration, the number of hidden layers can be determined first. As mentioned by Heaton (2008), one hidden layer can approximate any function that contains a continuous mapping from one finite space to another. With the increase of hidden layers, the NN model could have an excellent fitting performance, but it may decrease the prediction capacity due to overfitting. Since the dataset applied in this work contains only around 2,500 training data points, one hidden layer should work well enough and minimize the risk of overfitting (Bishop 2006). In this case, the training problem (equation (18)) can be rewritten as,

$$\hat{\mathbf{y}} = \sigma_2\left(\mathbf{W}^2\sigma_1\left(\mathbf{W}^1\mathbf{x}\right)\right) \quad (22)$$

Where, the activation function for the output layer σ_2 and hidden layer σ_1 is chosen as a linear function and ReLU, respectively.

The determination of weight matrixes is usually based on a back-propagation algorithm**.** In terms of equation (22), the partial derivative for the loss function *L* to $\mathbf{W}^1$ can be expressed as,

$$\frac{\partial L}{\partial \mathbf{W}^1} = \frac{\partial L}{\partial \hat{\mathbf{y}}}\frac{\partial \hat{\mathbf{y}}}{\partial f_2}\frac{\partial f_2}{\partial \mathbf{W}^2 f_1}\frac{\partial f_1}{\partial \mathbf{W}^1\mathbf{x}}\frac{\partial \mathbf{W}^1\mathbf{x}}{\partial \mathbf{W}^1} = \mathbf{W}^{2^T}\frac{\partial L}{\partial \hat{\mathbf{y}}}\frac{\partial f_1}{\partial \mathbf{W}^1\mathbf{x}}\mathbf{x}^T \quad (23)$$

$$\frac{\partial L}{\partial W^1_{m,j}} = \mathbf{W}^2_{m,}\frac{\partial L}{\partial \hat{\mathbf{y}}}\frac{\partial f_1}{\partial \mathbf{W}^1\mathbf{x}}\bigg|_{m,}\mathbf{x}_{j,}, \text{ where } \frac{\partial L}{\partial \hat{y}_{k,i}} = \lambda_k^2\left(\hat{y}_{k,i}-\hat{y}_{k,i}\right) \quad (24)$$

where $m \in M$ and M is the number of neurons in the hidden layer.

When $k=1$, namely, the model has a single output, each value in the linear transition matrix $W^1_{m,j}$ is only influenced by the single output. However, when $k>1$, $W^1_{m,j}$ are influenced by all outputs. Therefore, the correlations among different output variables could be incorporated during the training process.

4.5 Model training

In addition to the number of hidden layers, there are three other main parameters concerning the training of NN models: 1). the number of neurons *M* in the hidden layer; 2). The epoch number: one epoch represents one forward pass and one backward pass of all the

training examples; 3). Batch size: the number of training examples in one forward/backward pass.

As suggested by (Heaton 2008), the number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer, which equals to 10 in this analysis. But in reality, considering that each dataset has different characteristics, 6 neuron numbers are selected as candidates including 10, 16, 32, 48, 64 and 80. As for the epoch number, when it is small, the NN model may not be well trained. When it is large, the NN model has the potential to be over-trained. Here, three epoch numbers are considered, 50, 100, and 150. In addition, early-stopping is incorporated in order to avoid overfitting. Batch size determines the number of data points that can be trained for each time. When it is small, the NN model has the potential to be over-trained. Here, three batch sizes (8, 32 and 64) are considered.

The optimal combination of these three parameters is determined by grid search based on a 5-fold cross-validation. For each fold, all training data are randomly divided into three groups, including 70% training data, 20% validation data, and 10% test data. The training data and validation data are used to train the model and the test data is used to check the out-of-sample performance of the NN model. Theoretically, the dataset should be divided so that the three partitions – training, validation, and test – are large enough that we have a reasonable expectation that they are all representative of the same distribution. Practically, the ratio selection is mainly based on the data amount. For example, when the number of total data points is very large, 50% of the total data can be selected as the test data. When the number of data points is very small, there is only train and validation data, and no test data at all. Because, in this case, it is better to use as many training data as possible to fully train the model. As for Iowa dataset, the number of total data points for rigid pavements is around 2,700, so 10% of the test data is selected to make sure the model can be fully trained. By using the 5-fold cross-validation, the model can be tested by 5 different sets of test data.

The mean squared error (MSE) of the test data is chosen as the model performance metric, which is the same as the loss function (equation (19)) for the training process. After a 5-fold cross-validation, the average MSE is considered to compare NN models with different structure parameters. The model with the lowest average MSE was selected as the representative NN model.

5. Results

In order to explore the influence of incorporating multi-output variables, and various weights for each output variable on the model prediction performance, four single-output models were trained for each variable (i.e. IRI, FAULT, LCRACK and TCRACK) along with ten multi-output models with different sets of weights. The first multi-output model (listed as NN0 in Table 9) uses a set of weights that are uniform, [1, 1, 1, 1]. The second model (NN1) uses a set of weights equal to those used in the Iowa DOT PCI equation (equation (5)), i.e., [10, 5, 4, 6]. The other eight weight sets (for the other eight models, NN2 – NN10) are shown in Table 9 and explore how disparity of weighing among the four variables impacts model results.

The training speed for single-output and multi-output models is related with the training parameters, including the number of neurons, and the epoch number and the batch number. When both types of models use the same parameters: Neuron=64, Epoch=50, Batch=32, the training time for single-output and multi-output models are 26.3s and 29.8s,

respectively (the training time is based on 5-fold cross-validation). The difference between these two models in terms of training speed is not large.

Tables 9 lists the optimal set of model parameters and the smallest MSEs found during the grid search for these 4 single-output and 10 multi-output models, respectively. For the multi-output models, the optimal NN structure is selected based on average MSE of PCI prediction across a 5-fold cross-validation. Three things are immediately notable from this table. First, we see that, the multi-output models have better prediction performances (smallest MSEs) for IRI, FAULT, and TCRACK (see values with an asterisk in Table 9) compared to the corresponding single variable models. Both types of models have similar prediction performance for LCRACK. The improvement in prediction performance is particularly notable for the prediction of FAULT (single variable model MSE = 1.069, model NN1 MSE for FAULT=0.778, a 27% reduction) and for the highly influential prediction of IRI (single variable model MSE = 0.044, NN1 MSE for IRI = 0.033, a 25% reduction).

Table 9 Optimal single-output model structure and prediction performance (first four rows) and Optimal multi-output model structure and prediction performance (last 10 rows)

		NN structure			**Mean Squared Error (MSE)**				
Name		**Neuron**	**Epoch**	**Batch**	**PCI**	**IRI**	**FAULT**	**LCRACK**	**TCRACK**
IRI		32	150	32	11.78	0.044			
FAULT		80	150	8			1.069		
LCRACK		80	50	64				**468.9**	
TCRACK		32	50	8					137.9
Name	**Weights**	**Neuron**	**Epoch**	**Batch**	**PCI**	**IRI**	**FAULT**	**LCRACK**	**TCRACK**
NN0	[1, 1, 1, 1]	48	150	36	10.25*	0.034	0.764	482.3	133.1
NN1	[10, 5, 4, 6]	64	50	32	10.24* (0.035)	0.033* (2e-5)	0.778 (4e-4)	491.3 (0.3)	132.4* (0.04)
NN2	[10, 1, 1, 1]	80	100	8	10.37	**0.033***	0.800	489.2	141.4
NN3	[50, 1, 1, 1]	48	150	8	10.53	**0.033***	0.859	511.6	153.6
NN4	[1, 10, 1, 1]	80	50	8	11.09	0.038	**0.703**	501.0	140.1
NN5	[1, 50, 1, 1]	64	150	8	12.72	0.051	**0.690***	519.1	151.7
NN6	[1, 1, 10, 1]	48	100	8	11.03	0.038	0.827	**468.1***	136.4
NN7	[1, 1, 50, 1]	80	150	8	12.62	0.047	0.926	**470.2***	158.1
NN8	[1, 1, 1, 10]	48	50	8	10.68	0.039	0.835	500.8	**132.1***
NN9	[1, 1, 1, 50]	80	100	8	12.04	0.052	0.947	540.5	**132.3***

*Smallest observed MSE for that variable

Secondly, and more importantly, from Table 9 it is clear that the PCI prediction performance associated with either multi-output model NN0 or NN1 (MSE = 10.24) is better for than for the PCI prediction derived from the four optimal single output models (MSE = 11.78). As mentioned in section 4.4, the training process for multi-output models incorporates the correlations among variables. This correlation could improve the prediction performance for a single variable and should improve the prediction of a multi-variable metric like PCI.

Finally, although the multi-output models can outperform single metric models in predicting PCI, because PCI is a composite performance metric, when it is optimal, it does not mean all other single performance metrics are optimal.

Due to the existence of uncertainty in a multi-fold validation, the values listed in Tables 9 could change if the same model was regenerated. In order to compare different models statistically, a bootstrap analysis was conducted to estimate the standard deviation for the average MSE obtained by 5-fold cross-validation. Because this uncertainty in MSE

derives solely from the random partitioning of training and validation data, all 10 multi-output models have the same source of uncertainty. Therefore, a bootstrap analysis was conducted only for the NN1 model. The bootstrap sample size was 100. The standard deviation in MSE for each variable is listed in Table 9 (i.e. values in parentheses). Using these standard deviation values, statistically significant differences among the lowest observed MSE values was tested using a two-sample t-test assuming equal variance and a confidence level of 95%. The MSE values highlighted in Table 9 (asterisk) for PCI, IRI, FAULT, LCRACK and TCRACK are the lowest values. When multiple values in a given column have an asterisk, they are not statistically different.

In terms of PCI prediction, normally, the NN1 model would be expected to have the best prediction performance since its weights resemble the PCI formula. However, both NN0 and NN1 models have the smallest MSE. This is because in the current PCI formula, the coefficients (weightings) for all four variables are similar in magnitude. Figure 2 shows the comparison between true PCI and predicted PCI based on out-of-sample data for NN1 model, which presents a good prediction performance.

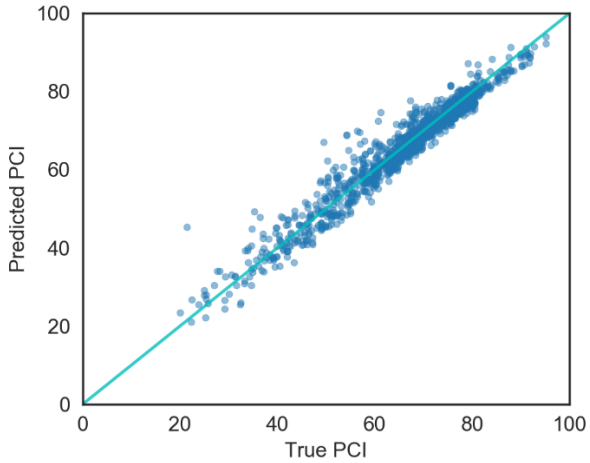
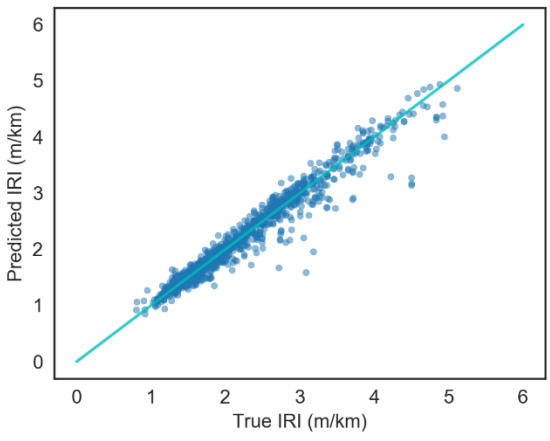


Figure 2. Comparison between true PCI and predicted PCI based on out-of-sample data when weights are [10, 5, 4, 6] (NN1 model).

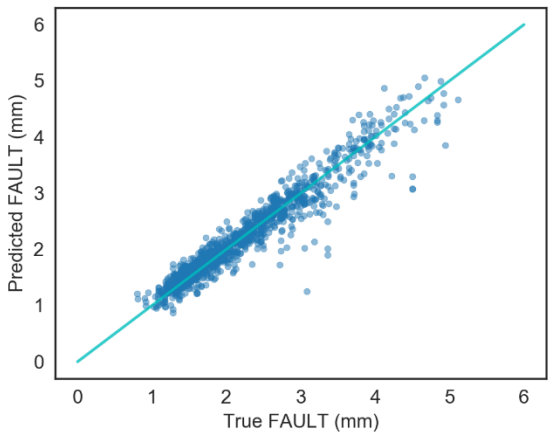
The models NN2 to NN9 have two larger weights (10x or 50x) for IRI, FAULT, LCRACK, TCRACK, respectively. For each of these models, the prediction performance of the variable with the largest weight (highlighted in bold) and the PCI prediction performance are emphasized. Generally, when an output is more heavily weighted, the training process emphasizes this output. As shown in equation (21), when the output Y_k is heavily weighted, the minimization process will concentrate more on improving the prediction performance on Y_k to minimize the total weighed loss value. When this occurs, prediction performance for that focal output would be expected to be higher than for an NN model with a smaller weight on that same output. It is important to note that models NN2 to NN9 are not expected to or intended to improve on the prediction of PCI compared to NN1. Instead the role of these models is to see how the multi-output model algorithm behaves as weighting shifts. In this case, where correlation is present, these also serve to demonstrate both the ability of a multi-output model to outperform single-output models and the fact that as weights become more

disparate there may be an important tradeoff between performance in predicting the composite metric (PCI) and the individual metrics.

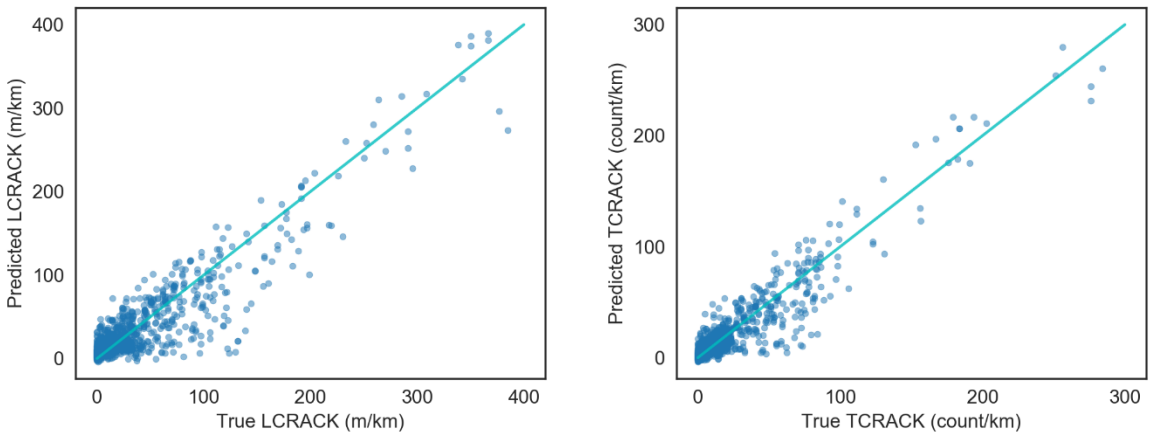
As for IRI, when its weight is 2 times larger than other variables (NN1), its MSE value is already the smallest. With an increase of IRI weight (models NN2 and NN3), IRI prediction performance stays almost the same while MSE for PCI prediction increases very slightly. The behavior of TCRACK prediction is similar to IRI. The NN1 model has a low MSE for TCRACK that is not statistically resolvable from models NN9 and NN10. However, TCRACK is influential in the prediction of PCI (only IRI is more influential). Therefore, as the TCRACK weight grows relative to the other variables (NN9 and NN10), PCI prediction performance deteriorates. For both FAULT and LCRACK, when they have larger weights (i.e. 10 and 50), their prediction performances are better than NN0 and NN1. Hence, with the increase of the weight for a single variable, its own prediction performance could be improved. However, other variables' prediction performances would decrease, as well as the PCI prediction performance. To this last point, it can be seen that the difference in model performance between a uniformly trained option and a weighted option increases as the heterogeneity of the weights grows. Figure 3 presents the comparison between true and predicted values for IRI, FAULT, LCRACK, and TCRACK. Generally, all four plots demonstrate good model performance without significant systematic deviations in correspondence between the predicted and actual values. Both LCRACK's and TCRACK's plots show a small number of low range values (~50 to 100 m/km LCRACK and ~50 to 100 count/km TCRACK) that are underestimated by the model. The authors believe this occurs for two reasons: (1). Data quality for these two variables is poor with standard deviations roughly two times the mean for the samples of both; (2). Some crack-related maintenance records may be missing. The record number of surface treatment (e.g. crack sealing) is extremely small. Hence, some input and output parings may not be appropriate when there exists a surface treatment but without a record. Underestimates at these low values should not cause significant problems at these are well below current treatment threshold triggers.



(a). IRI



(b). FAULT



(c). LCRACK (d). TCRACK

Figure 3. Comparison between true and predicted values based on out-of-sample data for (a). IRI (NN3 model); (b). FAULT (NN5 model); (c). LCRACK (NN7 model); (d). TCRACK (NN9 model)

6. Conclusions

A new weighted multi-output neutral network model for the deterioration of rigid pavements has been proposed based on the Iowa pavement management system database. This new model simultaneously predicts IRI, faulting (FAULT), longitudinal cracking (LCRACK) and transverse cracking (TCRACK) for pavements, providing convenience for pavement management systems whose treatment decisions are based on composite, multi-condition metrics such as the pavement condition index (PCI). Considering the fact that different condition metrics may have different importance levels, each metric to be predicted is given a weight during the model training process.

Compared to single-output models, an appropriately weighted, multi-output model (NN1) performs better at estimating PCI (13% lower MSE) than an estimate derived from four optimal, single-output models. Furthermore, the multi-output model generates better estimates than corresponding single-output models for three of the four individual metrics considered – IRI (25% lower MSE), FAULT (27% lower) and TCRACK (4% lower). For the fourth metric, LCRACK, the multi-output model has a 5% higher MSE than the single-output model. The observed improvements derive directly from incorporating correlation relationships among different variables within the multi-output model. The deterioration in LCRACK prediction performance demonstrates the tradeoff that is made when simultaneously estimating correlated metrics.

The results presented in this paper make it clear, that multi-output models can improve prediction performance in cases where correlation exists. Furthermore, variable weighting is important to achieve the optimal balance of prediction performance among the various metrics. Analysis of extremely weighted models (NN2 to NN9) suggests that the relevance of understanding the nature of this tradeoff will grow as weights become more heterogeneous.

As for the future work to improve the prediction performance for pavement deterioration, one aspect is to improve data quality, which is a common problem for data-driven deterioration models. The Iowa PMS provides one of the best pavement databases that

are available. However, there still exist some problems, such as large measurement errors, missing maintenance history, and data entry errors. On one hand, these problems increase the difficulty to generate the data for training the model, on the other hand, a large ratio of data points has to be removed. To improve the model performance, it is necessary to obtain data with a good quality. Another aspect is about the time series analysis. Currently, it is assumed that the time step is just two years due to limited data. With the increase of available records, it is necessary to test different time steps. As for the model structure, one potential drawback for neural network model is its 'black box' feature, which means that the relationship between input and output variables are not clear. In some research fields such as choice models in transportation, some attempts have been made to integrate theoretical model and data-driven models. In the future, this idea could also be applied in pavement deterioration analysis.

Acknowledgements

This research was conducted as part of the Concrete Sustainability Hub at the Massachusetts Institute of Technology, which is supported by the Portland Cement Association and the Ready Mixed Concrete Research and Education Foundation.

Declarations of interest: none

References

Abaza, K.A., 2004. Deterministic performance prediction model for rehabilitation and management of flexible pavement. *International Journal of Pavement Engineering*, 5 (2), 111–121.

Abaza, K.A., 2016. Back-calculation of transition probabilities for Markovian-based pavement performance prediction models. *International Journal of Pavement Engineering*, 17 (3), 253–264.

Aguiar-Moya, J.P., Prozzi, J.A., and de Fortier Smit, A., 2011. Mechanistic-empirical IRI model accounting for potential bias. *Journal of Transportation Engineering*, 137 (5), 297–304.

American Association of State Highway and Transportation Officials, 1993. *AASHTO guide for design of pavement structures*. Washington, DC.

American Association of State Highway and Transportation Officials, 2008. *Mechanistic-empirical pavement design guide. A manual of practice*. Washington, D.C.

Attoh-Okine, N.O., Mensah, S., and Nawaiseh, M., 2003. A new technique for using multivariate adaptive regression splines (MARS) in pavement roughness prediction. *Proceedings of the Institution of Civil Engineers - Transport*, 156 (1), 51–55.

Baker, D.C., Sipe, N.G., and Gleeson, B.J., 2006. Performance-based planning: perspectives from the United States, Australia, and New Zealand. *Journal of Planning Education and Research*, 25 (4), 396–409.

Bektas, Fatih; Smadi, Omar G.; and Al-Zoubi, M., 2014. *Pavement management performance modeling: evaluating the existing PCI equations*.

Ben-Akiva, M. and Ramaswamy, R., 1993. An approach for predicting latent infrastructure facility deterioration. *Transportation Science*, 27 (2), 174–193.

Bianchini, A. and Bandini, P., 2010. Prediction of pavement performance through neuro-fuzzy reasoning. *Computer-Aided Civil and Infrastructure Engineering*, 25 (1), 39–54.

Bishop, C.M., 2006. *Pattern recognition and machine learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Chen, X., Dong, Q., Gu, X., and Mao, Q., 2019. Bayesian analysis of pavement maintenance failure probability with Markov chain Monte Carlo simulation. *Journal of Transportation Engineering, Part B: Pavements*, 145 (2), 04019001.

Choi, J.H., Adams, T.M., and Bahia, H.U., 2004. Pavement roughness modeling using back-propagation neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 19 (4), 295–303.

Chu, C.Y. and Durango-Cohen, P.L., 2008. Estimation of dynamic performance models for transportation infrastructure using panel data. *Transportation Research Part B: Methodological*, 42 (1), 57–81.

FHWA, 2017. Highway Statistics 2017 [online]. Available from: https://www.fhwa.dot.gov/policyinformation/statistics/2017/ [Accessed 13 Aug 2019].

FHWA, 2019. LTPP InfoPave [online]. Available from: https://infopave.fhwa.dot.gov/ [Accessed 13 Aug 2019].

Gong, H., Sun, Y., Hu, W., and Huang, B., 2019. Neural networks for fatigue cracking prediction using outputs from pavement mechanistic-empirical design. *International Journal of Pavement Engineering*.

Heaton, J., 2008. *Introduction to neural networks for Java, 2nd Edition*.

Heidari, M.J., Najafi, A., and Alavi, S., 2018. Pavement deterioration modeling for forest roads based on logistic regression and artificial neural networks. *Croatian Journal of Forest Engineering*, 39 (2), 271–287.

Highway Research Board, 1962. *The AASHO Road Test, Special Reports No. 61-AE.* Washington, D.C.

Hong, F. and Prozzi, J.A., 2006. Estimation of pavement performance deterioration using Bayesian approach. *Journal of Infrastructure Systems*, 12 (2), 77–86.

Hong, F. and Prozzi, J.A., 2010. Roughness model accounting for heterogeneity based on in-service pavement performance data. *Journal of Transportation Engineering*, 136 (3), 205–213.

Hong, H.P. and Wang, S.S., 2003. Stochastic modeling of pavement performance. *International Journal of Pavement Engineering*, 4 (4), 235–243.

Hossain, M.I., Gopisetti, L.S.P., and Miah, M.S., 2019. International roughness index prediction of flexible pavements using neural networks. *Journal of Stomatology*, 145 (1).

Inkoom, S., Sobanjo, J., Barbu, A., and Niu, X., 2019. Prediction of the crack condition of highway pavements using machine learning models. *Structure and Infrastructure Engineering*, 15 (7), 940–953.

Iowa DOT Open Data [online], 2019. Available from: https://data.iowadot.gov/datasets/ [Accessed 30 Jul 2019].

Kargah-Ostadi, N., Stoffels, S.M., and Tabatabaee, N., 2010. Network-level pavement roughness prediction model for rehabilitation recommendations. *Transportation Research Record: Journal of the Transportation Research Board*, 2155, 124–133.

Karlaftis, A.G. and Badr, A., 2015. Predicting asphalt pavement crack initiation following rehabilitation treatments. *Transportation Research Part C: Emerging Technologies*, 55, 510–517.

Kobayashi, K., Kaito, K., and Lethanh, N., 2012. A statistical deterioration forecasting method using hidden Markov model for infrastructure management. *Transportation Research Part B: Methodological*, 46 (4), 544–561.

Lamptey, G., Labi, S., and Li, Z., 2008. Decision support for optimal scheduling of highway pavement preventive maintenance within resurfacing cycle. *Decision Support Systems*, 46 (1), 376–387.

Lee, B.J. and Lee, H.D., 2004. Position-invariant neural network for digital pavement crack analysis. *Computer-Aided Civil and Infrastructure Engineering*, 19 (2), 105–118.

Lethanh, N. and Adey, B.T., 2013. Use of exponential hidden Markov models for modelling pavement deterioration. *International Journal of Pavement Engineering*, 14 (7), 645–654.

Lou, Z., Gunaratne, M., Lu, J.J., and Dietrich, B., 2001. Application of neural network model to forecast short-term pavement crack condition: Florida case study. *Journal of Infrastructure Systems*, 7 (4), 166–171.

Mahmood, M., Rahman, M., and Mathavan, S., 2018. A multi-input deterioration-prediction model for asphalt road networks. *In*: *Proceedings of the Institution of Civil Engineers - Transport*. 1–12.

Mazari, M. and Rodriguez, D.D., 2016. Prediction of pavement roughness using a hybrid gene expression programming-neural network technique. *Journal of Traffic and Transportation Engineering (English Edition)*, 3 (5), 448–455.

Nii O. Attoh-Okine, 1994. Predicting roughness progression in flexible pavements using artificial neural networks. *3rd International Conference on Managing Pavements*, (2).

Owusu-Ababio, S., 1998. Effect of neural network topology on flexible pavement cracking prediction. *Computer-Aided Civil and Infrastructure Engineering*, 13 (5), 349–355.

Pan, N.F., Ko, C.H., Yang, M. Der, and Hsu, K.C., 2011. Pavement performance prediction through fuzzy regression. *Expert Systems with Applications*, 38 (8), 10010–10017.

Press, G., 2016. Cleaning big data: most time-consuming, least enjoyable data science task, survey says. *Forbes*, 23 Mar.

Prozzi, J.A. and Madanat, S., 2004. Development of pavement performance models by

combining experimental and field data sets. *Journal of Infrastructure Systems*, 10 (1), 9–22.

Prozzi, J.A. and Madanat, S.M., 2003. Incremental nonlinear model for predicting pavement serviceability. *Journal of Transportation Engineering*, 129 (6), 635–641.

Rice, J.A., 2006. *Mathematical Statistics and Data Analysis, 3rd edition*.

Roberts, C.A. and Attoh-Okine, N.O., 1998. A comparative analysis of two artificial neural networks using pavement performance prediction. *Computer-Aided Civil and Infrastructure Engineering*, 13 (5), 339–348.

Rosa, F.D., Liu, L., Asce, M., Gharaibeh, N.G., and Asce, M., 2017. IRI prediction model for use in network-level pavement management systems. *Journal of Transportation Engineering, Part B: Pavements*, 143 (1), 1–8.

Swei, O., Gregory, J., and Kirchain, R., 2018. Does pavement degradation follow a random walk with drift? Evidence from variance ratio tests for pavement roughness. *Journal of Infrastructure Systems*, 24 (4), 04018027.

Thube, D.T., 2012. Artificial neural network (ANN) based pavement deterioration models for low volume roads in India. *International Journal of Pavement Research and Technology*, 5 (2), 115–120.

Yao, L., Dong, Q., Jiang, J., and Ni, F., 2019. Establishment of prediction models of asphalt pavement performance based on a novel data calibration method and neural network. *Transportation Research Record*, 2673 (1), 66–82.

Yu, J., Chou, E.Y., and Luo, Z., 2007. Development of linear mixed effects models for predicting individual pavement conditions. *Journal of Transportation Engineering*, 133 (6), 347–354.

Zhang, W. and Durango-Cohen, P.L., 2013. Explaining heterogeneity in pavement deterioration: clusterwise linear regression model. *Journal of Infrastructure Systems*, 20 (2), 04014005.