Development of regression equations for local calibration of rutting and IRI as predicted by the MEPDG models for flexible pavements using Ontario's long-term PMS data

Gulfam-E-Jannat\textsuperscript{a}, Xian-Xun Yuan\textsuperscript{b} & Medhat Shehata\textsuperscript{b}
\textsuperscript{a} Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1
\textsuperscript{b} Department of Civil Engineering, Ryerson University, 350 Victoria Street, Toronto, ON, Canada, M5B 2K3
Published online: 31 Oct 2014.

To cite this article: Gulfam-E-Jannat, Xian-Xun Yuan & Medhat Shehata (2014): Development of regression equations for local calibration of rutting and IRI as predicted by the MEPDG models for flexible pavements using Ontario's long-term PMS data, International Journal of Pavement Engineering, DOI: 10.1080/10298436.2014.973024

To link to this article: http://dx.doi.org/10.1080/10298436.2014.973024

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Development of regression equations for local calibration of rutting and IRI as predicted by the MEPDG models for flexible pavements using Ontario’s long-term PMS data

Gulfam-E-Jannat\textsuperscript{a}\textsuperscript{*}, Xian-Xun Yuan\textsuperscript{b1} and Medhat Shehata\textsuperscript{b2}

\textsuperscript{a}Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1; \textsuperscript{b}Department of Civil Engineering, Ryerson University, 350 Victoria Street, Toronto, ON, Canada, M5B 2K3

\textsuperscript{*}Corresponding author. Email: gjannat@uwaterloo.ca

\textcopyright 2014 Taylor & Francis

http://dx.doi.org/10.1080/10298436.2014.973024

Local calibration is an important step before a transportation agency adopts the American Association of State Highway and Transportation Officials’ (AASHTO) mechanistic-empirical pavement design guide (MEPDG). This paper presents the challenges of and findings from the local calibration of flexible pavements in provincial highways under the jurisdiction of the Ministry of Transportation of Ontario (MTO). A calibration database was developed that involved a hierarchical framework of the input parameters required for AASHTOWare Pavement ME (the MEPDG software) and the historical field performance data based on the MTO’s second-generation pavement management system. A regression analysis is carried out for preliminary calibration of rutting and international roughness index (IRI) models by comparing the predicted distress to observed distress. The analysis suggested that whereas the MEPDG provided fairly unbiased prediction of the IRI value, it often over-predicted the total rutting. Calibrated predicted IRI and rut depth are found for Ontario’s local conditions from MEPDG distress prediction models. A further clustering analysis based on Functional Class and geographical zone for the rutting and IRI, respectively, improved the precision of the locally calibrated models.

Keywords: mechanistic-empirical; calibration; validation; rutting; international roughness index

1. Introduction

After many years of intense research and development, the MEPDG was finally proposed in 2008 for an interim design guide by American Association of State Highway and Transportation Officials’ (AASHTO 2008). An important component that plays a central role in the MEPDG is the empirical transfer models that relates the structural responses (i.e. stresses and strains) calculated in the mechanistic analysis to manifested pavement distresses (rutting and various cracking modes) and performance (measured by international roughness index, IRI). Under multiple National Cooperative Highway Research Program (NCHRP) projects including 1–37A (major development), 1–40 (software enhancement and model recalibration) and 9–30A (material characterisation and rutting model recalibration), a lot of efforts have been dedicated to calibrate and validate the transfer functions using pavement test sections throughout the North America, mainly based on the Long-Term Pavement Performance (LTPP) database. However, the empirical nature of the transfer functions dictates the needs of local calibration before a state or provincial transportation agency adopts the MEPDG for routine design. This step is important because the local policies and conditions often differ from the inference space determined by the global calibration database in terms of climate, traffic patterns, material selection, construction methods and maintenance practices. These differences may have significant impact on the predictive accuracy of the transfer models in the MEPDG. The local calibration is expected to mitigate the potential biases and reduce the variation of performance prediction (AASHTO 2010).

According to a recent survey (TRB 2009) as of the late 2009, 62% of the states in the USA had an implementation plan in place for the MEPDG. As one of the leading transportation agencies in North America, the Ministry of Transportation of Ontario (MTO) is also considering moving to the MEPDG protocol for future projects. This paper presents the challenges of and results from local calibration of the MEPDG by using data from the MTO’s pavement management system (PMS). The study will focus on flexible pavements, which account for a vast majority of the pavement network in Ontario. The objective of this study is to develop regression equations to preliminarily calibrate the MEPDG distress prediction models (for rutting and IRI) for Ontario’s local conditions.

2. Literature review

During the past few years since the AASHTO MEPDG was made available for trial use in 2004, there have been so many local calibration studies that a comprehensive literature review is difficult, if not impossible, in a reasonable space. Roughly speaking, the calibration...
studies have undergone the following three stages: (1) evaluation and validation, (2) local calibration by using limited local LTPP database or road test sections and (3) local calibration by using PMS data. A general conclusion that can be drawn from the studies in the first two stages was that local calibration was necessary in order to eliminate bias of the MEPDG distress prediction. The following review will mainly focus on the development in the third stage.

The use of PMS data for local calibration of MEPDG transfer models was a relatively new practice. Two different approaches have been seen in the literature: one is the so-called focus or clustering calibration, and the other is the section-by-section calibration. Some researchers also called the former Level 3 calibration and the latter Level 1 calibration. The focus calibration was first reported by Schram and Abdelrahman (2006, 2010) who analysed non-interstate flexible pavement sections under the jurisdiction of Nebraska Department of Roads. They divided the sections into 18 groups based on subgrade modulus and asphalt mix type. An interesting point brought up by the authors is the attempt to extend the prediction function of MEPDG for pavement management decisions. Obviously, this attempt can only be realistic after the MEPDG is accurately calibrated for local practice. In Canada, He et al. (2011) conducted a similar Level 3 calibration for the MEPDG rutting models based on AASHTOWare Pavement ME by using the long-term field data (about 20 years) from Alberta’s PMS (He et al. 2011). Instead of checking the rut depth at a project/section level, they adopted a network-level approach. Specifically, the inventory sections were divided into three categories, 14 groups depending on various pavement characterisation factors including rehabilitation, pavement materials and structures, and traffic volume. This grouping strategy was based on the hypothesis that the rutting pattern in these three categories should be different. ME analyses were then performed using group-averaged parameters in terms of pavement structures, performance, traffic and climate characteristics. From the study, they concluded that DARWin-ME (previous version of MEPDG softwares) over-predicts rut depth of new sections with non-stabilised granular base, whereas sections after rehabilitation involving milling are often moderately under-predicted. But DARWin-ME provides fairly close prediction for total rutting of sections rehabilitated by straight overlays with no milling.

There have been a few studies of section-by-section calibration using the PMS data. Li et al. (2009) calibrated the MEPDG distress models for new flexible pavement with data obtained from the Washington State PMS (Li et al. 2009). However, they used only 2 representative sections for local calibration and 13 other sections (two of them from the LTPP used for global calibration) for validation. The need of local calibration was concluded based on observable biases in various distresses between the predicted and observed values.

Using Iowa DOT’s comprehensive database, Kim et al. (2010) selected 16 representative pavement sections (both flexible and rigid) to evaluate the MEPDG performance predictions (Kim et al. 2010). Overall, local calibration of the MEPDG performance models was recommended to improve the accuracy of predictions. For rutting prediction in particular, MEPDG was found to systematically over-predict rut depth in hot mix asphalt (HMA) and under-predict rehabilitated sections of HMA over jointed plain concrete pavement. Large scatter was found in rutting prediction for HMA over HMA. In the study, they found a challenging issue in local calibration the definition and/or measurement discrepancy in HMA alligator and thermal cracking between MEPDG and Iowa’s PMS. This concern was shared in our study.

Hoegh et al. (2010) presented an interesting paper on local calibration of MEPDG rutting model based on the Minnesota Road Research Project (MnRoad) test sections. It was observed that the MEPDG over-predicted the rut depth for all the MnRoad pavement sections. Further investigation showed that the base and subgrade rutting prediction at early pavement age were consistently unrealistic high, even though all of the MnRoad pavement sections under examination were new HMA sections without overlay or other rehabilitations. Based on those observations, the authors proposed a non-conventional local calibration model that deducted the predicted rutting of the first month in the base and subgrade layers.

Using the online MEPDG software, Hall et al. (2011) reported results from their initial local calibration of flexible pavement models based on data from 18 LTPP sections and 20 PMS sections in Arkansas. They concluded that data availability and quality are critical for local calibration. In general, it was found that predicted distresses did not match well with measured ones. Although MEPDG caught the average trend for alligator and longitudinal cracking as well as IRI, large scatter was observed for the three distress/performance modes even after local calibration. Rut depth, however, was over-predicted by the nationally calibrated model. Transverse cracking prediction was actually even worse: MEPDG predicts no thermal cracking in all of the tested sections; however, transverse cracking was recorded in field distress surveys. The authors attributed this difference to definition discrepancy for transverse cracking between the MEPDG and LTPP.

Kim et al. (2011) developed a local calibration database for North Carolina flexible pavements that include 46 sections (22 LTPP sections for calibration and 24 PMS sections for validation). Preliminary local calibration and validation were carried out by using those sections. The results showed that the rut depth and fatigue cracking predictions were significantly different from the measured values. Different than the observations
from other states, the rut depth predicted by MEPDG was slightly under-estimated. The local calibration of the MEPDG was found to reduce bias and standard error of rut depth and alligator cracking. However, the improvement was not enough to accept the null hypothesis that the measured values were equal to the predicted values at the 95% confidence interval. The calibration results demonstrated the importance of using material-specific performance test results, having detailed and reliable distress data, and taking permanent deformation measurements from individual layers through forensic investigation. The study was performed in a hybrid platform with MEPDG version 1.1 and version 9–30A.

The possible interaction of the PMS and MEPDG was reviewed by Hudson et al. (2008) at a pretty abstract level. They concluded that it should be feasible to use the PMS data for MEPDG local calibration provided that an LTPP-type ‘Satellite PMS’ database is in place for long-term calibration. Therefore, it has been clear that development of a high-quality local calibration database is the key to the success of the calibration using PMS data. As mentioned, the LTPP database was used for global calibration, and it was designed to collect research quality pavement performance data. LTPP has sections of a fixed length of 500 ft, and distress evaluation is also standardised. In contrast, most of the states’ PMS data were collected mainly to support network-level decision-making. Sections in PMS are of varying length, and its distress evaluation methods are also varying with different agencies and with the LTPP practices. Corley-Lay et al. (2010) carried out a very interesting study in which the alligator cracking and rut depth measured by the LTPP and North Carolina’s DOT’s (NCDOT) PMS were compared. Their study showed significant differences exist between the LTPP-measured fatigue cracking and the percentage estimates obtained from the NCDOT windshield survey method. The LTPP walking survey captured more distress than did the windshield survey. Similarly, the NCDOT profile meter rut depth is consistently lower than the LTPP wire line rut depths. The low rut depth measurements may partly explain why DARWin-ME often over-predicts the rut depth in the previous local calibration studied reviewed above. Clearly, development of a high-quality calibration database is a very challenging and resource-intensive task.

## 3. Local calibration methodology

AASHTO published a guide for the local calibration of the MEPDG (AASHTO 2010), which will be referred to hereafter as the Guide. The Guide lists a step-by-step procedure for local calibration and discusses various issues in local calibration such as selection of accuracy levels of inputs, determination of sample size, selection of pavement sections, and specific procedure of eliminating bias and reducing standard errors if local calibration is needed. Following the spirit of the Guide, this study adopts a five-step local calibration process, namely (1) pavement database development, (2) evaluation of MEPDG, (3) local calibration by regression analysis, (4) clustering analysis and (5) validation.

The whole idea of local calibration is to compare the pavement distresses and performance predicted by the MEPDG against the actually observed values. For this purpose, a local calibration database needs to be developed at first. The database should include (1) observed distresses and performance data, and (2) input data for AASHTOWare Pavement ME analysis, the latter further grouped into the following categories: project information, traffic data, climate data, and pavement structure and material properties. The MTO has since the 1980s been using the PMS, which now has evolved into its second generation, referred to hereafter as the PMS2. Thus, the long-term field performance data from the PMS2 is used for developing the local calibration database. The database development involves the following important tasks:

- selection of pavement section-cycles;
- review and screening of historical performance data of selected sections and
- retrieve, review and screening of construction and rehabilitation documents for as-built information.

More details of the database development are discussed in Section 4.

AASHTOWare Pavement ME version 2.0 and the associated default transfer models are used in this study for pavement analysis and performance prediction. A unique feature of the local calibration study is that a fairly large number of pavement sections with long-term performance history recorded are selected for the analysis. This allows for a good number of design factors to be examined in the ME design. On the other hand, it also requires a great deal of efforts and computing time for pavement analyses in AASHTOWare Pavement ME. For this reason, this study does not adopt the conventional approach for eliminating bias and reducing standard error by determining the optimal values of the local coefficients in transfer models. Rather, regression analyses between the predicted and observed distresses and IRI values are performed. Admittedly, this is a pragmatic approach as well as a time-saving measure. An advantage of this approach over the conventional parameter-adjusting calibration is that the regression can always bring any potential biases to zero. Moreover, if the standard deviations (SDs) remain to be large after bias elimination, then a further regression analysis on the effects of a list of potential explanatory variables (e.g. surface material, overlay structure, traffic, subgrade strength and environmental zone) can be conducted, resulting in a clustering calibration model. After this is done, the calibrated models are then validated.
with another separate set of pavement sections. Thus, a split-sample approach is used for the local calibration-validation process.

4. Development of local calibration database

4.1 The MTO PMS2 performance data

The MTO oversees the maintenance and operation of 16,500 centre-line kilometres consisting of freeways, collectors, and arterial and local roads, of which 95% are bituminous pavements (Li and Kazmierowski 2008). As of the end of 2010, the historical pavement condition data stored in the MTO PMS2 database reached more than 48,000 section-years for the about 1800 pavement sections under the MTO jurisdiction. Unlike the standard length of 500 ft (approximately 152.5 m) used in LTPP, the Section Length in PMS2 ranges from 0.4 km to about 20 km, as the segmentation in PMS is largely based on project history, pavement structure/materials, traffic and environmental factors and so on.

The MTO follows a pavement performance condition rating practice that is not completely the same as per the AASHTO standards or LTPP data collection protocols (Miller and Bellinger 2003). Each year, conditions of all pavement sections (100% survey) are assessed in terms of roughness, rutting and distresses. Pavement roughness before 1997 was measured as riding comfort index (RCI) by use of Portable Universal Roughness Device. Since 1997, the Ministry has adopted IRI to measure the riding quality. Based on the IRI value, the RCI is reported through a correlation equation. Rutting measurement in Ontario was started in 2002, and it has been carried out by the Ministry staff using the Automatic Road Analyzer.

Surface distresses are evaluated by specially trained Pavement Design and Evaluation Officers using wind-shield-based method (Chong et al. 1989). For flexible pavement, 15 types of surface distress varying from ravelling, flushing, wheel-path rutting and distortion to various types of cracking are evaluated in terms of extent and severity at a scale from 0 to 5. These scores are then aggregated to a distress manifestation index (DMI) value ranging from 0 to 10 with 10 representing the perfect condition. Finally, a pavement condition index (PCI) that combines the RCI and DMI through a Ministry-prescribed equation is reported to evaluate the overall pavement condition and support maintenance prioritisation at a network level (Li et al. 2001).

One of the main challenges faced in the study is the conversion of the PMS performance data into a MEPDG compatible format. The distresses and performance predicted in the MEPDG for AC pavements are listed in Table 1 and a comparison of the units is shown. The IRI and total rut depth are already consistent, except that the Section Length might introduce additional statistical variation. But the cracking data are a major concern. The MEPDG considers fatigue cracking and thermal cracking separately. The fatigue cracking is further divided into two different modes (bottom-up and top-down) depending on whether the cracking is initiated by the tensile strain at the bottom of AC layer or by the shear strain at the edge of wheel load. For rehabilitated pavements, it also considers total cracking that includes reflective and alligator cracking. In the MTO’s surface distress evaluation, however, although various cracking patterns (longitudinal, transverse, alligator and random cracks) at wheel track, centre-line and pavement edge are separately evaluated, those evaluation data cannot be directly used for the calibration of the MEPDG predictions. It is often deemed that the alligator cracking and longitudinal cracking at wheel path be classified as bottom-up and top-down fatigue cracking, respectively, and that transverse cracking be attributed to thermal cracking. However, there were studies arguing that transverse and longitudinal cracking at wheel path may also be attributed to fatigue cracking as they can be just the initial stage of an alligator crack (Xiao et al. 2010). Moreover, as mentioned above, the cracking in the PMS2 are assessed in scores. Even though the scores (often of the extent rather than the severity) could be converted back to a numeric value, the units used for scoring the longitudinal and transverse cracking in PMS2 are different than those in the MEPDG, as shown in Table 1.

Therefore, only the terminal IRI, total rutting and AC bottom-up fatigue cracking are the candidate distresses, which transfer models, may be locally calibrated in this study. For the other distresses, further study on the conversion system is needed before local calibration can be carried out.

Table 1. Comparison of units of predicted and measured distresses.

<table>
<thead>
<tr>
<th>Distress type</th>
<th>Unit in MEPDG</th>
<th>Unit in MTO PMS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminal IRI</td>
<td>m/km</td>
<td>m/km</td>
</tr>
<tr>
<td>Permanent deformation – total pavement</td>
<td>mm</td>
<td>mm</td>
</tr>
<tr>
<td>Total cracking (reflective + alligator)</td>
<td>Percentage</td>
<td>Not recorded</td>
</tr>
<tr>
<td>Asphalt concrete (AC) bottom-up fatigue cracking</td>
<td>Percentage</td>
<td>Rating based on percentage</td>
</tr>
<tr>
<td>AC top-down fatigue cracking</td>
<td>m/km</td>
<td>%</td>
</tr>
<tr>
<td>Permanent deformation – AC only</td>
<td>mm</td>
<td>Not recorded</td>
</tr>
</tbody>
</table>
4.2 Selection of the pavement sections and life cycles

A guiding principle for the selection of pavement sections is efficiency of local calibration, which means a balance between the efforts of data collection and benefits from local calibration in terms of bias elimination and precision enhancement. Two questions need to be answered: (1) how many test pavement sections (i.e. sample size) should be used, and (2) which sections should be selected? Based on consideration of statistical significance and variability of distresses and IRI in LTPP database, the Guide suggests minimum numbers of sections for each distress type (AASHTO 2010). The Guide also recommends the use of fractional factorial design to pick up test sections. In this study, the following factors are mainly considered in the selection of pavement sections:

- pavement performance curve of a section cycle (complete cycle between the treatments);
- availability and quality of historical performance data;
- pavement types and rehabilitation strategies;
- region and transportation corridor;
- facility types (freeway, arterials, collectors and local) and
- asphalt materials.

For most of the pavement sections, the long-term performance data spans over several life cycles of structural rehabilitation or reconstruction. Because the construction records in the PMS2 are unreliable and incomplete, a set of heuristic rules were developed to identify the years of structural rehabilitation and reconstruction based on the condition data. The sections of a full life cycle with quality historical performance data are selected for candidate. Because of the long performance history in PMS2, some of the selected sections had two life cycles.

All of the three types of flexible pavement (conventional, deep strength and full depth) were designed in Ontario. A majority of the current pavement sections in PMS2 are rehabilitated or reconstructed. In addition, a great variety of rehabilitation treatments have been used in Ontario, including hot-mix overlay, milling before overlay, reclamtion or recycling before overlay, and so on (Li and Kazmierowski 2008). As for asphalt materials, both Marshall and Superpave mixes are used.

Finally, 96 section cycles from 82 sections are selected for the analysis. The pavement features considered by the database are summarised in more detail in Table 2. It is shown that a wide range of values for a great number of parameters are covered by the selected sections. For the calibration–validation purposes, the data are split into a calibration set of 68 section cycles and a validation set of 28 section cycles.

<table>
<thead>
<tr>
<th>Features</th>
<th>Number or range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section cycles</td>
<td>96</td>
</tr>
<tr>
<td>No. of sections</td>
<td>82</td>
</tr>
<tr>
<td>Pavement types</td>
<td>Conventional (new – 2;</td>
</tr>
<tr>
<td></td>
<td>rehabilitated – 73;</td>
</tr>
<tr>
<td></td>
<td>reconstruction – 15);</td>
</tr>
<tr>
<td></td>
<td>deep strength – 6</td>
</tr>
<tr>
<td>Highway routes</td>
<td>23</td>
</tr>
<tr>
<td>Functional Classes</td>
<td>Freeways and arterials</td>
</tr>
<tr>
<td>HMA types</td>
<td>8 (mainly Marshall mixes)</td>
</tr>
<tr>
<td>Environmental zones</td>
<td>4 (southern, centre, northwestern and northeastern)</td>
</tr>
<tr>
<td>Subgrade soils</td>
<td>4 (mainly sandy silt)</td>
</tr>
<tr>
<td>Subgrade modulus</td>
<td>15–70 MPa</td>
</tr>
<tr>
<td>AADT</td>
<td>6000–140,000</td>
</tr>
<tr>
<td>Length of life cycle</td>
<td>6–18 years</td>
</tr>
<tr>
<td>Section Length</td>
<td>1.8–19.6 km</td>
</tr>
<tr>
<td>Section width</td>
<td>7–12.4 m</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>2, 3 or 4</td>
</tr>
</tbody>
</table>

4.3 Accuracy levels of input data

After the sections and associated life cycles are selected, the remaining task of database development is to collect the design inputs for each section cycle. This was done by reviewing contract documents and as-built records. A major issue with this task is the accuracy levels of the inputs.

MEPDG adopts a three-level accuracy hierarchy for the input variables including climate, traffic loading, layer structure and material properties. Generally speaking, Level 1 input is the most accurate value that is possible to obtain for a certain input. This usually represents site-specific measurements and laboratory data. On the other extreme, Level 3 inputs are either the default values used in AASHTOWare Pavement ME or nominal values in a local agency. In global calibration and recalibration under NCHRP projects 1–37A and 1–40D, the highest accuracy levels in the LTPP database were used. For local calibration, however, the Guide seemed to suggest using the same accuracy levels for local calibration as the agency would use for future design. This practice is understandable for the purpose of evaluating the MEPDG performance for existing pavements, but it may be inefficient because the benefits from extra effort for obtaining more accurate input data for local calibration could easily outweigh the efforts made for a single design.

In this study, great efforts have been devoted to ensure the best quality data being collected. Table 3 summarises the accuracy levels used in the local calibration database. The least accurate inputs were related to material properties. There is an ongoing project in Ontario to characterise the dynamic modulus of the asphalt materials.
5. Results and discussions

5.1 Assessment of the MEPDG global models

The 68 calibration pavement section cycles of the database developed from MTO PMS2 were analysed using AASHTOWare Pavement ME with the default transfer models and calibration parameters. For comparison purposes, the 50% reliability predictions at the end of life of each section are used. As mentioned previously, only the alligator cracking, total rut depth and IRI are available in the historical performance data.

Figure 1(a)–(c) shows the scatter plots of the predicted versus the observed values of the three variables. As shown in Figure 1(a), both the predicted and observed fatigue cracking of most of the sections in the database are centred at zero. (To better visualise the repeated data points, small random perturbations are added to both observed and predicted values in the figure) Only a few sections of 35% observed cracking are predicted to zero fatigue damage. There are also two sections being over-predicted by the MEPDG. With these data, further local calibration is impossible. Nevertheless, a further investigation is needed before a confirmative conclusion can be drawn for the AC bottom-up fatigue cracking.

From Figure 1(b), it is observed that for a majority of the sections the predicted total rut depths are much higher than the field observed values. This observation is in agreement with many previous local calibration studies in other states, as reviewed in Section 2. In some states, the rutting in granular layers and subgrade was even forced to be zero in order to obtain an unbiased rutting prediction.

Table 3. Accuracy levels of AASHTOWare Pavement ME inputs used in the database.

```
<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project information</td>
<td>All information is project-specific.</td>
<td>Total rutting for pre-existing conditions is assumed 4 mm when unknown</td>
</tr>
<tr>
<td>Climate</td>
<td>Water table height (6.1 m for all sections).</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The closest climate station is selected from</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the list for distance &lt; 10 km; otherwise a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual Station is created</td>
<td></td>
</tr>
<tr>
<td>Traffic</td>
<td>Axle geometry, tire configurations, axle</td>
<td>Monthly adjustment factor, hourly distribution factor</td>
</tr>
<tr>
<td></td>
<td>distribution</td>
<td></td>
</tr>
<tr>
<td>Materials/structures</td>
<td>Asphalt binder properties</td>
<td>Dynamic modulus, creep compliance and other thermal parameters</td>
</tr>
<tr>
<td></td>
<td>(Ontario Provincial Standard values for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>different region are used)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dynamic modulus, creep properties</td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 1. Comparison of predicted with observed distresses: (a) bottom-up (alligator) cracking, (b) rutting and (c) IRI.
In this study, unrealistically large rut depth was obtained for those new, long life-cycle sections. This suggests that the traffic exponent parameter $\beta_{2.3}$ for AC rutting model should be $<1.0$. For rehabilitated sections, in addition to the five local calibration parameters commonly used for bias elimination, there are two other inputs of Level 3 accuracy that are found to be very sensitive to the rutting prediction. These are the pavement condition rating and the pre-existing total rutting before rehabilitation. As shown in Table 3, for pre-existing condition, a 4 mm total rut depth before rehabilitation was used when it is unavailable. The existing pavement condition rating directly affects the stiffness calculation of the damaged structural layers, whereas the pre-rehabilitation total rut depth affects the incremental rut depth calculation on the existing layers. A simple linear regression between the observed and predicted rut depths yields the following empirical Equation (1) and a simple linear regression between the observed and predicted values yields the following empirical Equation (2) for IRI.

\[
RD_{\text{calibrated predicted}} = 5.3167 + 0.1058 RD_{\text{model predicted}},
\]

with $R^2 = 0.0785$ only and residual SD = 1.775 mm.

\[
IRI_{\text{calibrated predicted}} = 0.3377 + 0.7790 IRI_{\text{model predicted}},
\]

with $R^2 = 0.3417$ and residual SD = 0.302 m/km.

The MEPDG provides better prediction for the IRI than for the previous two distresses; in fact, only slight prediction on average is observed in Figure 1(c). This is also in agreement with observations from previous studies in other states. The enhanced performance of the IRI model may be attributed to its empirical nature in MEPDG.

Another comparison was made for the failure prediction. In real life, among the 68 sections selected for calibration analysis, 14 sections have not yet reached their end of life cycle, whereas the other 54 had been rehabilitated after their life cycle. This means that 54 of 68 sections have ‘failed’ by a certain criterion dictated by DMI, RCI and PCI, as it was the practice in MTO. However, the AASHTOWare Pavement ME predicts that only 10 sections would have failed by whichever of the ME criteria used in MEDPG for flexible pavements. The implication is interesting: even though the rutting is over-predicted, the MEPDG would have failed to recommend rehabilitation for the MTO for many of sections. This discrepancy might be attributed to the pre-set failure thresholds chosen for MEPDG by the Ministry. A further study is clearly needed in this regard.

### 5.2 Clustering and validation

To improve the accuracy and precision for the IRI and rutting models, clustering regression analyses are carried out. The following factors are considered as candidate clustering variables:

- pavement type (conventional and deep strength; new, rehabilitated and reconstructed);
- highway Functional Class (freeways and arterials);
- geographical zone;
- subgrade modulus;
- annual average daily traffic (AADT) and
- Section Length.

The Section Length is included in the list because it might affect the variation of the observed rut depth and IRI. The inclusion of the other factors is obvious as they are the common influential parameters for pavement performance.

A stepwise regression is first performed to remove all possible high-order interaction terms in the regression analysis. To facilitate the selection of the clustering variable the $R^2$ values of the regression models for the calibration data, and residual SDs of both the calibration and validation sections are calculated. For the validation data, the bias of each candidate model is also calculated. To assess the accuracy, the $p$-values for the biases based on a two-sided $t$-test with the alternative hypothesis of ‘Bias $\neq 0$’ are reported. To avoid over-fitting, it is usually required that the validation SD be not greater than the calibration SD. For this purpose, the $p$-values for the one-sided $F$-test are also reported. The results are summarised in Tables 4 and 5 for rutting and IRI, respectively. As shown in Table 4, the goodness of fit in terms of $R^2$ is much improved by introducing one or two clustering variables. The $p$-values show no bias of the clustered regression model for the validation data. Also the validation SD for all candidate models does not show significantly large than the calibration SD. In fact, except for the case of subgrade modulus, all the validation SDs are lesser than the calibration SD. Moreover, when only one parameter is used, the Functional Class is the best clustering variable with highest $R^2$ value and lowest calibration SD and validation SD. When two clustering variables are used, the stepwise regression (results not shown here due to space limitation) suggests that the Section Length is the best alternative. As shown in the table, with this additional variable, the $R^2$ value is further improved to 0.388 and calibration SD further reduced. This suggests that the Section Length may have significant impact on rutting prediction indeed. Figure 2(a) further illustrates the improvement of local calibration for the rutting model. Without clustering, the simple regression can indeed bring the bias to zero. However, the calibrated predictions are scattered widely around the 45° line. The
clustering regression by using the Function Class and Section Length further improve the precision of the prediction. The situation for IRI is slightly different. As shown in Table 5, among the five clustering variables, the Functional Class actually provides the least $R^2$ value.

The zone and subgrade modulus have the same $R^2$ value of 0.49, but when the subgrade modulus is used for clustering the validation SD is much higher than the calibration SD. Thus, overall geographical zone has the best clustering performance. That is, the local calibration models for IRI should be separated for different administration regions.

A figurative comparison of the IRI results before and after clustering analysis is illustrated in Figure 2(b). Similar to the case for rutting, after clustering analysis, more data are concentrated around the 45° line.

Although, a better correlation is found between the MEPDG-predicted and observed IRI, further investigations are required to confirm the precise predictions of all distresses before going to next step of calibrating the transfer functions.

Table 4. Regression analysis results for rut depth.

<table>
<thead>
<tr>
<th>Clustering variables</th>
<th>Degrees of freedom</th>
<th>No clustering</th>
<th>Pavement type</th>
<th>Functional Class</th>
<th>AADT</th>
<th>Zone</th>
<th>Subgrade modulus</th>
<th>Functional Class + Section Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD (calibration)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD (validation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bias (validation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$-Value for bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$-Value for SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Regression analysis results for IRI.

<table>
<thead>
<tr>
<th>Clustering variables</th>
<th>Degrees of freedom</th>
<th>No clustering</th>
<th>Pavement type</th>
<th>Functional Class</th>
<th>AADT</th>
<th>Zone</th>
<th>Subgrade modulus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD (calibration set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD (validation set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bias (validation set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$-Value for bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$-Value for SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Regression analysis results for rut depth.

clustering regression by using the Function Class and Section Length further improve the precision of the prediction. The situation for IRI is slightly different. As shown in Table 5, among the five clustering variables, the Functional Class actually provides the least $R^2$ value. The zone and subgrade modulus have the same $R^2$ value of 0.49, but when the subgrade modulus is used for clustering the validation SD is much higher than the calibration SD. Thus, overall geographical zone has the best clustering performance. That is, the local calibration models for IRI should be separated for different administration regions. A figurative comparison of the IRI results before and after clustering analysis is illustrated in Figure 2(b). Similar to the case for rutting, after clustering analysis, more data are concentrated around the 45° line.

Although, a better correlation is found between the MEPDG-predicted and observed IRI, further investigations are required to confirm the precise predictions of all distresses before going to next step of calibrating the transfer functions.

![Figure 2](image-url)  
**Figure 2.** Comparison of the calibrated predictions with observations before and after clustering regression: (a) rut depth; (b) IRI.
6. Conclusions

Results from the first large-scale local calibration in Ontario that used long-term field performance data from PMS database were reported in this paper. Great efforts have been devoted to the selection of pavement sections and collection of design input data for the use of AASHTOWare Pavement ME analysis. Because Ontario uses different pavement surface distress evaluation approach, most of the historical cracking data in the PMS2 could not be directly used. Thus, only total rut depth and IRI are mainly investigated in this study. The analyses showed that the MEPDG over-predicted rutting for majority of the road sections, whereas the IRI default model had better performance on average. The analysis suggested that most of the Ontario’s roads were resilient to fatigue cracking, but this finding needs further investigation for a confirmative conclusion.

To improve accuracy and precision, this study also performed a clustering regression analysis using the calibration and validation sections. It was found that the highway Functional Class was a good clustering variable for rutting, whereas for IRI, the geographical zone (which is in parallel with MTO’s administration regions) served better. Moreover, the Section Length was found to help explain the variation in the observed rutting values.

Due to computational constraints, this study did not calculate the optimal local calibration coefficients in the transfer models. Rather, regression analyses between the predicted and observed distresses and IRI were conducted. However, the conventional local calibration by taking the former approach is under the way. In addition, several challenging issues that are remained to be addressed for further study include

- collection of Level 1 material properties for asphalt mixes;
- conversion and classification of various PMS2 cracking data into the three cracking types predicted by MEPDG and
- calibration of the AASHTOWare Pavement ME in terms of life prediction.

Acknowledgements

The authors would like to thank Dr Ningyuan Li, Dr Joseph Ponniah and Mr Warren Lee of Pavement and Foundation Section of the MTO for their help and insightful discussions during the course of the research work.

Funding

This research was funded by a grant from the MTO under the Highway Infrastructure Innovation Funding Program. The financial support of MTO is highly appreciated.

Notes

1. Email: arnold.yuan@ryerson.ca
2. Email: mshehata@ryerson.ca

References


