Empirical Comparison of Statistical Pavement Performance Models

Chih-Yuan Chu[1] and Pablo L. Durango-Cohen[2]

Abstract: We conduct an empirical comparison of nine representative statistical pavement performance models, which have been estimated using serviceability data from the AASHO Road Test. The purpose of the study is to understand the effect of different statistical assumptions and estimation techniques on the models' predictive capabilities. The study consists of using the models to predict the serviceability of a common subset of pavements from the test over a two-year forecast horizon. Comparison of the models is carried out using both aggregate- and disaggregate-level measures. The main insight that stems from our study is that models that account for heterogeneity lead to improved predictive capabilities. This is particularly important because the data were obtained from a controlled experiment. The study further shows that disaggregate measures are more useful in testing performance models because they are better indicators of the ability to capture the variability in data. Throughout the paper, we also discuss other model characteristics and issues that can be useful in the future development and evaluation of performance models.

DOI: 10.1061/(ASCE)1076-0342(2008)14:2(138)

CE Database subject headings: Infrastructure; Pavements; Models; Predictions; Comparative studies.

Introduction

A key component in making design, construction, maintenance, and rehabilitation decisions for transportation infrastructure, e.g., pavements, consists of evaluating the effect of these decisions on the performance of such facilities. The evaluation, in turn, involves assessing and measuring surface distresses (e.g., cracking and rutting) or structural properties (e.g., deflection and strain) and forecasting the effect of the aforementioned decisions on future conditions. Condition forecasts are generated with performance models, which are mathematical expressions that relate condition data to a set of explanatory variables such as design characteristics, traffic loading, environmental factors, and the history of maintenance activities.

The scale of expenditures associated with the above managerial decisions, as well as the far-reaching and serious negative economic and social impact of deficient infrastructure have, over the last 40 years, motivated a great deal of research in the development of statistical performance models. Numerous estimation techniques have been employed to estimate these models under different statistical assumptions, using countless data sources, and to address and support a gamut of purposes and managerial decisions. The capabilities and features of these models have been analyzed and evaluated qualitatively in the literature (McNeil et al. 1992; Gendreau and Soriano 1998); however, a quantitative comparison has, for the most part, been lacking. Madanat et al. (2002) and Li (2005) are, perhaps, notable exceptions, although their focus is different from the writers. Thus, the objectives of the present study are:

1. To provide a framework based on objective and quantitative measures that can be used (in certain situations) to evaluate and select statistical performance models; and
2. To understand the effects of various characteristics (i.e., statistical assumptions and estimation techniques) on the predictive capabilities of representative, state-of-the-art performance models.

To achieve the above objectives, we compare the predictive capabilities of nine representative, state-of-the-art models that were developed to forecast the "serviceability," i.e., the ride quality or functional performance, of pavement sections from the American Association of State Highway Officials (AASHO) Road Test (Highway Research Board 1962). The comparison is based on both disaggregate/section-level metrics. These metrics are related to the "pavement management levels" as defined by the AASHTO Joint Task Force on Pavements (AASHTO JTFP 2001). Our choice of data set and performance models allow for a meaningful and rigorous comparison, and are justified because:

- The AASHO Road Test provides high-quality, detailed pavement data, as well as weather and traffic records. The data set has been adopted by a plethora of studies using various methodologies and the estimation results have been reported for decades; and
- Furthermore, pavement design standards in the United States (and elsewhere) are largely based on models estimated with the functional performance data collected in the AASHO Road Test (AASHTO 1993).

It is important to recognize that, ultimately, other choices of data sources, performance models, and measures exist. Therefore, the detailed results and conclusions of our analysis may not gen-

[1]Assistant Professor, Dept. of Transportation Technology and Supply Chain Management, Kainan Univ., S209, No.1 Kainan Rd., Luzhu Shiang, Taoyuan 33857, Taiwan. E-mail: jameschu@mail.knu.edu.tw

[2]Assistant Professor, Dept. of Civil and Environmental Engineering and Transportation Center, Northwestern Univ., 2145 Sheridan Rd., A335, Evanston, IL 60208 (corresponding author). E-mail: pdc@northwestern.edu

Note. Discussion open until November 1, 2008. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on November 29, 2006; approved on September 24, 2007; published online ahead of print on March 4, 2008. This paper is part of the ***Journal of Infrastructure Systems***, Vol. 14, No. 2, June 1, 2008. ©ASCE, ISSN 1076-0342/2008/2-138–149/$25.00.

Table 1. Experimental Design: AASHO Road Test

Factor/loop	Factorial test section	Traffic loading Lane 1 (kN)	Traffic loading Lane 2 (kN)	Surface thickness (cm)	Base thickness (cm)	Subbase thickness (cm)
1	49	0	0	2.5, 7.6, 12.7	0, 15.2	0, 20.3, 40.6
2	44	8.9S	26.7S	2.5, 5.1, 7.6	0, 7.6, 15.2	0, 10.2
3	60	53.4S	106.8T	5.1, 7.6, 10.2	0, 7.6, 15.2	0, 10.2, 20.3
4	60	80.1S	142.3T	7.6, 10.2, 12.7	0, 7.6, 15.2	10.2, 20.3, 30.5
5	60	99.6S	177.9T	7.6, 10.2, 12.7	7.6, 15.2, 22.9	10.2, 20.3, 30.5
6	60	133.4S	213.5T	10.2, 12.7, 15.2	7.6, 15.2, 22.9	20.3, 30.5, 40.6

Note: S=single axle loading and T=tandem axle loading.

eralize to other situations. Another important point to note is that the main subjects of comparison are the statistical models. One must be careful when extending the conclusions of this research to the methodologies and estimation techniques that are used to estimate these models. Indeed, for the same methodology, the results of estimation are heavily dependent on a modeler's preference, model specifications, estimation routines, data transformation, and/or data preprocessing. Nevertheless, we believe that the work described herein provides valuable insights that apply to the development and evaluation of statistical performance models.

In the remainder of the paper we briefly review the salient features of the models that we use in our study. In particular, we describe the statistical assumptions that are used to formulate each of the models, as well as the estimation procedures that were employed. We then present the results of an empirical study where we compare the predictive capabilities of nine pavement performance models. We begin the following section, however, by providing background information about the AASHO Road Test, including an overview of the experimental design.

Data Set: AASHO Road Test

The AASHO Road Test was conducted between October 1958 and November 1960 near Ottawa, Ill., about 80 miles southwest of Chicago. The site was chosen because the soil is uniform and representative of the soils in a large area of the United States and because the climate is typical of those in the northern United States. The objective of the test was to collect research data useful for establishing the relationships between (flexible and rigid) pavement performance and important factors such as pavement structure design (surface, base, and subbase thicknesses) and traffic loading (axle applications, load, and configuration). In this paper, we focus on statistical performance models developed for flexible pavements.

The AASHO Road Test is one of the most comprehensive full-scale accelerated pavement tests conducted in the United States. The goal of the experimental design was to observe pavement deterioration through failure, and study only a few, most important variables. The comprehensive test of these important factors, including traffic loads and pavement structure configurations, and the care with which other factors were controlled explain why the data collected during this experiment are still one of the most widely used sources in the development of pavement performance models and pavement design criteria. In the remainder of this section, we provide an overview of the experimental design.

The test tracks in the AASHO Road Test consisted of six loops with two lanes each. Each loop was constructed as a succession of pavement sections. Each pavement section was further separated into two lanes by the centerline of the pavement section. The minimum length of a section was 4.6 m (15 ft) in Loop 1 and 30.5 m (100 ft) in others. The two lanes of Loop 1 were not subjected to traffic for the purpose of testing the effect of environmental factors. The other 10 lanes had approximately the same number of axle applications during the two-year period. However, the axle load and configuration were different for each lane. In addition to traffic loading, the sections were built with different combinations of surface, base, and subbase thicknesses. Each loop had at least one pavement section for every combination of traffic, surface, base, and subbase thicknesses. Loops 3–6, for example, had two traffic levels and three thicknesses for each layer, which totals $2\times3\times3\times3=54$ combinations. The other six sections in the loops were replicates. The traffic loading and structural design are summarized in Table 1.

The surface distress measurements collected through the duration of the study included rut depth, slope variance, cracking, and patching. Inspections were carried out biweekly over the project's duration. Thus, 56 observations were available for the sections that remained functional throughout the study. Most studies that have used this data set either focused on modeling the progression of one of the individual distresses, or of a functional performance index that aggregates the effect of multiple distresses. In this paper, we consider statistical models that were formulated so that the dependent variable corresponds to the present serviceability index (PSI), which was developed as part of the AASHO Road Test. The PSI is a widely accepted indicator of a pavement's functional performance and is computed as follows:

$$\mathrm{PSI} = 5.03 - 1.91\log_{10}(1+\mathrm{SV}) - 1.38\,\mathrm{RD}^2 - 0.03\cdot\sqrt{K+P} \tag{1}$$

where SV=*slope variance* (in.^{2}); RD=*rut depth* (in.); and K and P correspond to cracking and patching (ft^2). The PSI is a continuous index with values ranging from 5 (the best) to 0 (the worst).

Performance Models

In this section we review the performance models considered in the empirical study, focusing on the description of the main statistical assumptions that are used in each of the formulations. To facilitate the presentation, we have tried to use the same conventions and notation. In particular, and where appropriate, each pavement section is identified with a subscript i. The index $t=1,2\ldots,56$ is used to order the inspections/observations within the sequence for each facility. We also use the index t to label the

J. Infrastruct. Syst. 2008.14:138-149.

Table 2. Model Characteristics

Model	Source	Latent performance	Heterogeneity	Incremental models	Serial dependence
AASHO	Highway Research Board (1962)	No	No	No	No
AASHO(T)	Small and Winston (1988)	No	No	No	No
AASHO(D)	Prozzi and Madanat (2000)	No	No	No	No
OLS	Prozzi and Madanat (2004)	No	No	Yes	No
RE	Prozzi and Madanat (2004)	No	Yes	Yes	No
IM	Chu and Durango-Cohen (2007)	Yes	Yes	Yes	Yes
SUTSE	Chu and Durango-Cohen (2007)	Yes	Yes	Yes	Yes
SE	Chu and Durango-Cohen (2007)	Yes	No	Yes	Yes
OP	Li (2005)	Yes	No[a]	No	No

[a]Random-effects ordered probit model for bridge–deck, which considers heterogeneity, has been proposed by Madanat et al. (1997).

periods between inspections. $T=56$ denotes the duration of the study.

We begin this section by summarizing the characteristics that we use to describe the nine models considered herein. Each model is identified in Table 2 by the source article where it was introduced. The model characteristics are as follows:

- Latent performance: Refers to the assumption of unobservable dependent variables. Models specified under this framework assume that $PSI_{i,t}$ is an (imperfect) indicator of pavement section i's unobservable serviceability/functional performance at time t.
- Heterogeneity: Refers to the assumption that pavements are different or heterogeneous, which is in contrast to the assumption that pavements are homogeneous and deteriorate identically under the same condition.
- Incremental models: Refer to models that are specified under the assumption that incremental changes in the explanatory variables cause incremental condition changes. This is in contrast to models specified under the assumption that condition is determined by the cumulative effect of the explanatory variables, e.g., traffic loading.
- Serial dependence: Refers to the assumption that the dependent variable, $PSI_{i,t}$, exhibits autocorrelation, meaning that it is (statistically) dependent on (a subset of) the prior sequence of dependent variables.

In the remainder of the section, we discuss the implications of the above characteristics in the context of the performance models considered in the study. We also use these characteristics to interpret the results of the empirical study presented in the next section.

Highway Research Board (1962)

The original AASHO model for flexible pavements is shown in Eqs. (2)–(4). As can be seen, the model is nonlinear, which motivated the writers to consider a two-stage linear regression approach for estimation. It is noteworthy that in Eq. (2) ρ can be interpreted as traffic applications to failure because when a section fails [i.e., $PSI(W)=PSI_1$], ρ equals the current traffic applications (W). Moreover, β can be interpreted as a rate of deterioration because W/ρ represents the ratio of current traffic repetitions to total traffic repetitions, and thus the loss of PSI is a function of W/ρ and β. We use the label AASHO to identify this model

$$PSI(W) = PSI_0 - (PSI_0 - PSI_1)\left(\frac{W}{\rho}\right)^{\beta} \quad (2)$$

$$\beta = 0.4 + \frac{0.081(L_1 + L_2)^{3.23}}{(SN + 1)^{5.19} L_2^{3.23}} \quad (3)$$

$$\rho = \frac{10^{5.93}(SN + 1)^{9.36} L_2^{4.33}}{(L_1 + L_2)^{4.79}} \quad (4)$$

where W=accumulated axle load applications at the time when the serviceability is observed; $PSI(W)$=PSI value given as a function of W; and PSI_0=Initial PSI value. For the test sections in the AASHO Road Test PSI_0 was set to 4.2. PSI_1=PSI value at which the sections are considered failed. In the AASHO Road Test, sections with average PSI below 1.5 were classified as failed. The average is taken over the inner and outer wheel paths. L_1=nominal load axle weight (kip); L_2=indicator variable where $L_2=1$ for single axle vehicles and $L_2=2$ for tandem axle vehicles; SN=structural number; $SN=0.44D_1+0.14D_2+0.11D_3$, where D_1, D_2, and D_3 are the thicknesses of surface, base, and subbase layer (in.); and β, ρ=model parameters requiring estimation. These are determined by the axle load and configuration, as well as by the structural design.

The fundamental assumptions in this model are that:

- Pavement condition (in a given period) is determined by the cumulative effect of traffic.
- Observations/measurements in different periods or for different values of the explanatory variables are independent, i.e., serial dependence/correlation is ignored. This means that pavement condition is assumed to be fully determined by exogenous variables, and explains why the subscript t is not necessary in the equations.

The AASHO model has been criticized by various researchers (Small and Winston 1988; Paterson 1987). The major critiques include poor fit to data, use of an inefficient sequential estimation procedure, mismatched units, and misspecified model. Nonetheless, and as noted earlier, pavement design standards in the United States and elsewhere, i.e., AASHTO (1993), are largely based on these equations. To account for subgrade types and environmental conditions that are different from those in the original study, the above equations are modified as shown in AASHTO (1993).

Small and Winston (1988)

As discussed in the previous subsection, the estimation procedure of the AASHO model is cumbersome. An implication of setting $W=\rho$ [and thus, $PSI(W)=PSI_1$] in Eq. (2), is that ρ corresponds to the number of axle applications that will cause a pavement section to fail. Small and Winston (1988) point out that the left-hand

J. Infrastruct. Syst. 2008.14:138-149.

side of Eq. (4) is observable, and thus, that the equation should be estimated directly. Because some of the pavement sections outlasted the experiment, which means that only the lower bounds on the part of ρ are observed, the equation is estimated using Tobit regression (or censored regression). The updated equation is shown in Eq. (5). In the empirical study, we use this new estimate of ρ [along with the previous estimate of β in Eq. (3)] to update the AASHO model [Eq. (2)]. The new model is labeled AASHO(T)

$$\rho = \frac{10^{5.24}(\mathrm{SN}+1)^{7.761}L_2^{3.238}}{(L_1+L_2)^{3.652}} \tag{5}$$

Prozzi and Madanat (2000)

The writers present a stochastic duration model where time to failures are assumed to follow a Weibull distribution, an assumption that is often used in structural reliability. Similar to Small and Winston (1988), the focus is to update the equation for the loadings to failure, ρ, while rigorously accounting for the censored observations that are present in the data. The result is converted to the form of the AASHO model and are shown below in Eq. (6). In the empirical study, we use the expected value of ρ, $E[\rho]$, to update the AASHO model and label the corresponding model AASHO(D)

$$E[\rho] = \frac{10^{5.28}(\mathrm{SN}+1)^{6.68}L_2^{2.62}}{(L_1+L_2)^{3.03}} \tag{6}$$

Prozzi and Madanat (2004)

This study contributes to the literature in two ways. First, this study adopts *incremental* models as a different approach to estimate pavement performance models using AASHO Road Test data. The premise in these models is that incremental loadings, along with other explanatory variables (e.g., structural design and the prevailing environmental conditions), cause incremental changes in condition. In addition to being intuitive from a physical perspective, the writers explain that incremental models are appealing because managerial decisions usually rely on incremental predictions for a few periods. Note that serial independence is still assumed, and thus, incremental predictions do not depend on current or prior conditions. The second contribution of this study is that panel data modeling is adopted to account for unobserved heterogeneity in the serviceability data. Unobserved heterogeneity refers to the presence of section-specific effects that affect deterioration and are unexplained by the exogenous variables, such as construction quality. If not accounted for, unobserved heterogeneity can lead to biased and inefficient parameter estimates (Frees 2004).

In the study, the writers use incremental nonlinear regression models. The models are specified as shown in Eqs. (7)–(9). Eq. (7) is for the PSI of a given facility in a given period, which is obtained by aggregating the incremental PSI changes. Eq. (8) represents the incremental change in PSI occurring in time period t, i.e., $\Delta\mathrm{PSI}_{i,t} \equiv \mathrm{PSI}_{i,t+1}\text{-}\mathrm{PSI}_{i,t}$. An interesting feature of Eq. (8) is that incremental PSI changes depend on the cumulative traffic loading and on the frost depth observed during the period. These dependencies are consistent with the notions of fatigue and freeze–thaw cycles, respectively, which are known to affect deterioration. The incremental PSI changes also depend on the incremental traffic loading, which is represented in Eq. (9). Essentially, the equation converts incremental traffic loads into equivalent standard loads. This is an extension to the idea of equivalent single axle loads (ESALs), which is widely used in pavement performance modeling to normalize the effect of traffic loadings

$$\mathrm{PSI}_{i,t} = \alpha_1 + u_i + \alpha_2 \exp\{\alpha_3 D_{1i}\} + \sum_{r=1}^{t-1} \Delta\mathrm{PSI}_{i,r} + \varepsilon_{i,t} \tag{7}$$

$$\Delta\mathrm{PSI}_{i,t} = (1 + \alpha_4 D_{1i} + \alpha_5 D_{2i} + \alpha_6 D_{3i})^{\alpha_7}[\exp\{\alpha_8 G_{t-1}\} N_{i,t}^{\alpha_9} \Delta N_{i,t-1}] \tag{8}$$

$$\Delta N_{i,t} = n_{i,t}\left[\left(\frac{\mathrm{FA}_i}{\alpha_{10}18}\right)^{\alpha_{12}} + d_i\left(\frac{\mathrm{SA}_i}{18}\right)^{\alpha_{12}} + d_i\left(\frac{\mathrm{TA}_i}{\alpha_{11}18}\right)^{\alpha_{12}}\right] \tag{9}$$

where $\mathrm{PSI}_{i,t}$=PSI value of section i at the start of period t; $\Delta\mathrm{PSI}_{i,t}$=incremental PSI change of section i during period t; D_{1i}, D_{2i}, D_{3i}=thickness of surface, base, and subbase layers for pavement section i (in.); d_i=indicator of axle configuration for section i (d_i=1 for one rear axle; d_i=2 for two rear axles); $n_{i,t}$=number of truck passes for section i during time period t; $N_{i,t}$=cumulative loading for section i up to time t; $\Delta N_{i,t}$=traffic in ESALs for section i during period t; G_t=frost gradient in period t (in/day); FA_i, SA_i, SA_i=load of the front axle, single axle with dual wheels, and tandem axle with dual wheels, respectively (kip); and u_i=section-specific random error term with $E[u_i]=0$ and $Var[u_i]=\sigma_u^2$. For the OLS model, $\sigma_u^2=0$. For the RE model, u_i is assumed randomly distributed in the population and σ_u^2 is estimated to be 0.126. The OLS and RE models will be defined shortly. $\varepsilon_{i,t}$=time and section-specific random error terms with $E[\varepsilon_{i,t}]=0$ and $Var[\varepsilon_{i,t}]=\sigma_\varepsilon^2$. The σ_ε^2 estimate is 0.062 for the OLS model and 0.142 for the RE model. (α_1, α_2, α_3, α_4, α_5, α_6, α_7, α_8, α_9, α_{10}, α_{11}, α_{12})=estimated parameters, which are (4.45, −1.47, −0.555, 2.28, 0.775, 0.546, −2.67, −0.186, −0.473, 0.790, 1.72, 3.57) for the OLS model and (4.24, −1.43, −0.856, 1.39, 0.329, 0.271, −3.03, −0.173, −0.512, 0.552, 1.85, 4.15) for the RE model.

Instances of the above model are estimated using two estimation approaches: ordinary least-squares (OLS) nonlinear regression and the random-effects (RE) model. Panel data models account for unobserved heterogeneity by assigning section-specific, linear error terms, u_i, which generally provide more robust parameter estimates. The interpretation of the error terms is that pavement sections have variable intercepts due to material density, drainage conditions, foundation type, etc. In the RE model, a class of panel data models, the error terms are assumed to be randomly drawn from a population whose error terms are assumed to be normally distributed. As a result, instead of estimating individual error terms for each facility, it is only necessary to estimate the second moments associated with the distribution of population of error terms, which reduces the number of parameters. Although individual random effects in the RE model can be estimated and used to improve the prediction (Reinsel 1984); these individual terms cannot be transferred to other situations without requiring updates, and thus its transferability would be low. Thus, the predictions in Prozzi and Madanat (2004) do not include these effects and represent the average behavior of the population. The advantage of doing so is that transferability increases while the disadvantage is that the improvement of prediction might not be observed. On the other hand, the OLS model is essentially a classical regression model and assumes that

J. Infrastruct. Syst. 2008.14:138-149.

unobserved heterogeneity does not exist. In other words, none of the parameters are section specific. The estimation results of the two models are listed along with the equations, and the PSI values at time t can be obtained using Eqs. (7)–(9).

Chu and Durango-Cohen (2007)

This study proposes state-space specifications of multivariate dynamic models as a framework to develop models for infrastructure facilities. The main assumption in dynamic models is that the condition data are serially dependent, meaning that a facility's condition in a given period depends on the sequence of prior observations/transitions, as well as on exogenous explanatory variables and random error terms. This is fundamentally different from the models described earlier, which are specified under the assumption that pavement condition (in a given period) is determined by the cumulative traffic loading (up to that point). Dynamic models also are beneficial from a pavement management perspective since incremental, short-term predictions can be easily made given the current condition. However, due to serial dependence incremental predictions of dynamic models are dependent on the current condition while those of incremental models (i.e., OLS and RE models) are not.

Three classes of models are considered based on different assumptions regarding the underlying mechanisms generating the data sequences. The models are as follows:

- Individual models (IMs): Specified under the assumption that the deterioration of the individual facilities are instances of different and independent stochastic processes. Essentially, this approach uses an individual dynamic model to represent the deterioration of each pavement section. The approach is impractical due to the large number of parameters requiring estimation. In addition, it is not possible to estimate the effect of variables that are constant throughout the experiment for a given section, e.g., SN. Finally, the estimation can be interpreted as characterizing a stochastic process from a single realization. In turn, this precludes formulating general inferences about the underlying process, and consequently, there is no basis for out-of-sample prediction; that is, using a model estimate with data from one facility to predict the performance of other facilities. We do, however, use this framework in the empirical study to benchmark the predictive capabilities of the other statistical performance models.
- Seemingly unrelated time series models (SUTSEs): Specified under the assumption that the deteriorations of the individual facilities are instances of different stochastic processes that exhibit common elements and are contemporaneously correlated. In particular, the model presented in the paper has common parameters for the autoregressive terms describing the effects of serial dependence, traffic ($TRF_{i,t}$), and maintenance ($OVR_{i,t}$), as well as section-specific parameters for random error terms, $\omega_{i,t}$. The section-specific random error terms are used to capture the heterogeneity of variances (heteroscedasticity). These terms provide a great deal of flexibility but complicate the estimation. Similar to the RE model, the properties of the parameter estimates of SUTSE would improve due to the consideration of heterogeneous variances. It is noted that better properties of parameter estimates are not equivalent to superior predictions since SUTSE still predicts the average behavior of pavement sections and individual heterogeneity is not used for prediction.
- Single equation models (SEs): Specified under the assumption that the deterioration of the individual facilities are instances of the same stochastic process. Since the sections have common elements, the number of parameters is small.

The final specifications and estimation results are presented in Eqs. (10)–(15).

IM

$$x_{i,t} = \theta_{1,i} x_{i,t-1} + h_{1,i}\mathrm{TRF}_{i,t-1} + h_{2,i}\mathrm{OVR}_{i,t-1} + \omega_{i,t} \tag{10}$$

$$10\mathrm{PSI}_{i,t} = x_{i,t} + \xi_{i,t} \tag{11}$$

SUTSE

$$x_{i,t} = 0.995 x_{i,t-1} - \mathrm{SN}_i^{-1.062}\mathrm{TRF}_{i,t-1} + 15.349\mathrm{OVR}_{i,t-1} + \omega_{i,t} \tag{12}$$

$$10\mathrm{PSI}_{i,t} = x_{i,t} + \xi_t, \xi_t \sim N(0, 1.071^2) \tag{13}$$

SE

$$x_{i,t} = 0.984 x_{i,t-1} + 0.067\mathrm{SN}_i - 0.207\mathrm{TRF}_{i,t-1} + 15.739\mathrm{OVR}_{i,t-1} + \omega_t,$$

$$\omega_t \sim N(0, 1.852^2) \tag{14}$$

$$10\mathrm{PSI}_{i,t} = x_{i,t} + \xi_t, \xi_t \sim N(0, 1.112^2) \tag{15}$$

where $x_{i,t}$=latent performance of section i at time t. A larger value indicates a better condition. $10\mathrm{PSI}_{i,t}$=pavement condition in 10PSI of section i at time t. $\theta_{1,i}$=first-order autoregressive parameter for section i. For the SUTSE and SE models, the parameters are replaced by the estimated values (0.995 for SUTSE and 0.984 for SE). The parameters for the IM model are too many and thus not listed due to space limitations. $\mathrm{TRF}_{i,t}$=seasonal-weighted equivalent single axles in 10^5 ESALs of section i during time t. The seasonal weighted function is defined by the Highway Research Board (1962). $\mathrm{OVR}_{i,t}$=1 if overlay is applied during time t and 0 otherwise. $h_{1,i}, h_{2,i}$=parameters associated with $\mathrm{TRF}_{i,t}$ and $\mathrm{OVR}_{i,t}$, respectively, for section i. The parameters for the IM model are too many and thus not listed due to space limitations. For the SUTSE and SE models, the parameters are replaced by the estimated values. $\omega_{i,t}$=normally distributed random error of deterioration process of section i at time t. If the sections have a common deterioration process error, index i is dropped. The numbers of parameters for SUTSE and IM are too large and thus not listed due to space limitations. $\xi_{i,t}$=normally distributed random error of measurement process of section i at time t. If the sections have a common measurement error, index i is dropped. The parameters for the IM model are too many and thus not listed due to space limitations.

The use of traffic loading in ESALs simplifies the estimation because it reduces the number of parameters in the model. However, as a result of using aggregate explanatory variables, the impact of axle loads and configurations cannot be extracted from the results. Similarly, SN is used to simplify the models/analysis.

The state-space modeling framework allows for the specification of simultaneous equations representing the deterioration and inspection processes. This explains why each of the model instances presented above consist of two (sets of) equations. The *deterioration* model relates the explanatory variables to a latent performance variable that can be interpreted as the true serviceability. The *measurement* model relates the latent performance variable to the PSI and captures random errors induced by the fact that the measurements/observations are imperfect manifestations of the underlying serviceability.

J. Infrastruct. Syst. 2008.14:138-149.

An attractive feature of dynamic models in state-space form is that intervention analysis can be used to estimate the effect of maintenance activities. This approach is used in the above models to estimate the (transitory) effect of overlays, $OVR_{i,t-1}$, on $PSI_{i,t}$.

Li (2005)

Madanat et al. (1995) were first to consider the ordered probit (OP) model in the context of infrastructure performance modeling. Li (2005) was first to estimate such models using data from the AASHO Road Test. The OP technique is used in situations where the dependent variables, i.e., condition data, are used to construct or are presented in the form of discrete and ordinal ratings/categories, e.g., very good, good, fair, poor, or very poor. The reasons for considering this approach using data from the AASHO Road Test that does not follow this structure include:

- The complexity and cost of measuring the distresses that are needed to calculate the PSI for in-service pavements motivate the need for a different approach to develop performance models. In particular, it is attractive to develop methodology that can be driven by condition data collected using methods such as subjective ratings provided by inspectors, which are practical for large networks. The AASHO Road Test provides an opportunity to validate such an approach with high-quality data.
- The level of detail contained in continuous indices such as the PSI often exceeds the need to support managerial decision making. In particular, a coarse characterization of a condition is usually sufficient to select appropriate interventions. This idea appears in maintenance and repair optimization models formulated as Markov decision processes (see, e.g., Golabi et al. 1982; Golabi and Shepard 1997). As shown by Madanat et al. (1995), the OP technique provides an approach to estimate the transition probabilities that are needed in such models, although the assumption of serial independence in OP constitutes the most critical limitation for pavement performance modeling (Mishalani and Madanat 2002). In addition, the use of a few discrete categories provides an appealing way to represent the distribution of a condition across a system of facilities. We make use of this idea when we consider aggregate performance prediction metrics in the empirical study presented in the following section.

The OP model builds on the latent performance framework described earlier. Discrete condition ratings are observed indicators of the latent deterioration. In Eq. (16), the unobserved condition of section i at time t, $U_{i,t}$, is related to exogenous variables such as pavement thickness, traffic, and climate. Note that only the discrete rating, $C_{i,t}$, is observable. The model assumes that when $U_{i,t}$ falls within ϕ_{4-k} and $\phi_{4-(k+1)}$, $C_{i,t}=k$ will be observed as Eq. (17) shows. Since ε_i is assumed to be the standard normal distribution, Eq. (18) is the probability that the predicted discrete rating is k. The parameters of explanatory variables, β_0, β_1, β_2, β_3, and β_4, and the thresholds, ϕ_1, ϕ_2, and ϕ_3, are obtained by maximizing the likelihood that observed and predicted ratings are equal for all facilities

$$U_{i,t} = \beta'\Theta_{i,t} + \varepsilon_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 \mathrm{CTRF}_{i,t} + \beta_4 S_{i,t} + \varepsilon_{i,t} \tag{16}$$

$$C_{i,t} = k \quad \text{if and only if} \quad \phi_{4-(k+1)} - \beta'\Theta_{i,t} < \varepsilon_i \leq \phi_{4-k} - \beta'\Theta_{i,t} \tag{17}$$

$$P(C_{i,t}=k) = \Phi(\phi_{4-k} - \beta'\Theta_{i,t}) - \Phi(\phi_{4-(k+1)} - \beta'\Theta_{i,t})$$

$$\forall k = 0,1,\ldots,4 \tag{18}$$

where $U_{i,t}$=latent deterioration of section i at time t: The latent variable here captures the deterioration of a pavement section, which is opposed to latent performance in IM, SUTSE, and SE that describes serviceability. Therefore, a smaller value indicates a better condition in the OP models. $C_{i,t}$=discrete PSI rating of section i at time t as defined in the previous section; $\Theta_{i,t}$=vector of explanatory variables of section i at time t, i.e., $(1, D_{1i}, D_{2i}, \mathrm{CTRF}_{i,t}, S_{i,t})'$; β=vector of parameters of explanatory variables, i.e. $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$; $\mathrm{CTRF}_{i,t}$=cumulative equivalent single axles in ESALs of section i up to time t; D_{1i}, D_{2i}=thicknesses of surface and base layers of section i (in.); $S_{i,t}$=spring season dummy variable of section i at time t, 1 for spring and 0 otherwise; ε_i=random error term of section i following standard normal distribution, that is, $N(0,1)$; and Φ=standard normal cumulative distribution function (CDF).

The estimated parameters and thresholds are as follows: $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)=(1.19379, -0.10967, -0.01859, 3.00657E^{-7}, 0.05729)$ and $(\phi_{-1}, \phi_0, \phi_1, \phi_2, \phi_3, \phi_4)=(-\infty, 0, 1.67585, 2.66006, 3.76197, \infty)$.

Note that the OP model predicts probabilities associated with each state. To compare with the single-value predictions of the other models, we calculate an "expected state" by considering the weighted sum of the state midpoint PSI value and the probability of being in that particular state. Thus, the predictions assume that the midpoint PSI value, e.g., 3.5 for a good state ($3 \leq \mathrm{PSI} < 4$), is representative of the condition of pavement sections in that particular state.

Empirical Comparison

In this section we present the results of an empirical comparison of the predictive capabilities of the performance models presented in the previous section. We first define the terminology and scope of the comparison and describe the data that were used to conduct the comparison. We then use the performance models to predict the evolution of a set of pavement sections from the AASHO Road Test over the two year duration of the experiment. Finally, we compare the predictions to the actual data using both disaggregate- and aggregate-level measures and present the results of our study.

In general, comparison of statistical models can be made at three levels: model estimation, in-sample prediction, and out-of-sample prediction. At the model estimation level, goodness-of-fit and the properties of parameter estimates are examined. For in-sample prediction, the models are used to reproduce the samples that are used for estimation. For out-of-sample prediction, the models are used to predict observations that are not used in the estimation.

In this paper, we compare the in-sample predictive capabilities of the models introduced in the previous section. The decision is justified because:

- Tests at the model estimation level, e.g., the properties of parameters, measures of goodness-of-fit, and model diagnosis, are presented in the original studies. In other words, the models reported in the original studies are statistically satisfactory. Further, we note that a comparison at this level would be rather limited (and qualitative in nature) due to the structural differences between the models, the statistical techniques, and

J. Infrastruct. Syst. 2008.14:138-149.

Table 3. Data Set for Comparison

Group	Subgroup	Number of sections	Average number of observations[a]	Used for estimation by[b]	Included in comparison
Loop 1 (no traffic)					
	Lane 1	24	56	IM, SUTSE, SE	No
	Lane 2	24	56	None	No
Late failure (complete record with overlay)					
	Before overlay	87	27	All	Yes
	After overlay	87	29	IM, SUTSE, SE	No
No failure (complete record without overlay)					
	—	77	56	All	Yes
Early failure (no overlay)					
	—	120	18	AASHO	No
	—	120	18	AASHO(T)	No
	—	120	18	AASHO(D)	No
	—	120	18	OLS, RE, OP	No
Total	—	332	—	—	—

[a]The average number of observations is taken with respect to the numbers of sections.

[b]In the original studies where OP, IM, SUTSE, and SE are presented, portions of the data are set aside to validate the models.

data used for estimation. As an example, it is rather meaningless to compare R^2 values in regression models with ρ^2 values in discrete choice models.

- While the ultimate objective of comparing pavement performance models would be to conduct out-of-sample predictive comparison, numerous factors, primarily related to the amount of data and the original experimental design, prevent us from carrying this out in an extensive and rigorous fashion. Out-of-sample comparison would, for example, involve re-estimating the above models with a portion of the data set and validating them with another portion. Another difficulty is that it would be hard to ensure consistency with the original studies. For example, numerical implementation details are seldom reported in the original studies, meaning that the algorithms used for estimation or the type and extent of data preprocessing are unknown.

Note that in-sample predictive comparison may seem trivial to compare regression models because model fitting at the model estimation level and in-sample prediction are essentially the same. In this regard, we note that the predictive comparison carried out herein only uses a subset of the data used in the estimation of each of the models, i.e., the data used for estimation and for comparison are different, which means that the results reported are different from those in the original study. We also note that in-sample prediction is a powerful testing tool for the dynamic models and for the OP model presented in the previous section. For the dynamic models, one-step (two-week) prediction is used for model estimation, and in-sample prediction involves up to a 56-step prediction for the two-year duration of the study. In the OP model, the model estimation maximizes the likelihood of observing the predicted state. In-sample predictions generate the entire state distribution from which we calculate the mean state and compare it with the corresponding observation. In summary, in-sample prediction is practical, reasonable, and provides a quantitative basis to compare across the nine models presented above. The actual analysis is described in the next subsections.

Selection of Comparison Data Set for the Empirical Study

The selection of a comparison data set is critical to ensure a level playing field, and thus, a meaningful comparison. The difficulty in selecting a data set is that, as a result of different statistical assumptions, the models presented in the previous section are equipped to cope with different types of data. These differences are best explained by considering the failure patterns that were observed during the experiment together with the measures that were taken to correct the failures. The discussion below, including the description of the data that were used in the empirical comparison, is summarized in Table 3.

Overall, the AASHO Road Test consisted of 332 main factorial design sections. Data for 308 pavement sections were available for the analysis because the raw data from Lane 2 of Loop 1 were missing. The sections exhibited three failure patterns. Note that sections with average PSI (taken over the inner and outer wheel paths) below 1.5 were classified as failed. Several sections were observed to have severe distresses in one of the wheel paths and they were considered failed even though the average PSI was above 1.5. The first pattern (no failure) was for sections that did not fail during the experiment. The second pattern (early failure) was for sections that failed early during the experiment. These sections were reconstructed after the failure and taken out of the test. That is, no distress measurements were collected after the reconstruction. Finally, the third pattern (late failure) was for sections that failed toward the end of the experiment. These sections were overlaid, reopened to traffic, and continued to be monitored. All of the performance models incorporated no failure data for estimation. As presented by Chu and Durango-Cohen (2007), dynamic models (SE, SUTSE, and IM) exclude early failure sections since they are not capable of dealing with unbalanced panel data; that is, when the number of observations for facilities in the panel is unequal. The other models include the data that preceded the failures for the sections exhibiting the early failure pattern. Dynamic models are able to cope with (and estimate the effect of)

J. Infrastruct. Syst. 2008.14:138-149.

Table 4. Disaggregate-Level Comparison

Model	RMSE (PSI)	Ranking	MAE (PSI)	Ranking	PEV (PSI^2)	Ranking	ME (PSI)	ME (PSI^2)	Ranking
AASHO	0.897	7	0.467	7	0.783	8	−0.145	0.0209	5
AASHO(T)	1.633	9	0.942	9	1.981	9	−0.829	0.6869	9
AASHO(D)	0.989	8	0.664	8	0.711	7	−0.516	0.2665	8
OLS	0.590	6	0.435	5	0.223	2	0.353	0.1247	7
RE	0.485	2	0.370	2	0.225	3	0.101	0.0103	2
IM	0.429	1	0.277	1	0.170	1	0.119	0.0142	4
SUTSE	0.541	3	0.378	3	0.239	4	0.232	0.0539	6
SE	0.551	4	0.405	4	0.293	5	−0.104	0.0109	3
OP	0.574	5	0.440	6	0.330	6	0.016	0.0003	1

Note: RMSE=root-mean-square error; MAE=mean absolute error; PEV=prediction error variance; and ME=mean error.

exogenous interventions. Thus, the observations following late failures and subsequent overlays are included in the estimation. The other models only use the data before the overlays, i.e., the data collected after the overlays are discarded. Another difference is that dynamic models (SE, SUTSE, and IM) include sections that are not subject to traffic, while the other models do not. In summary, the data set used for comparison included the observations from the pavement sections that outlasted the test without overlay (no failure) and those from the pavement sections overlaid prior to the time of overlay (late failure–before overlay).

In the original studies where OP, IM, SUTSE, and SE are presented, portions of the data are set aside to validate the models. Such data, however, are included in the comparison results reported in the paper. The reason for including these data is that a random sampling is used in Li (2005) to select observations for validation and this information is not available. We believe that this approach is still acceptable since, in essence, no extrapolation (out-of-sample prediction) has been done. For the OP model, the ranges of variables of the comparison data set remain the same as those of the estimation data set. For the IM, the samples that are not used for estimation are the replicates of samples in the estimation data set. We understand this approach is not in favor of these models; however, it is very different from the situation where, for example, the models are estimated using the samples prior to overlay and compared using the samples after overlay.

Disaggregate/Section-Level Comparison

Disaggregate-level prediction and planning are used to develop effective management strategies for selected pavement sections within available funds and other constraints (AASHTO JTFP 2001). The comparison measures that we used at the disaggregate-level are root-mean-squared error (RMSE), mean absolute error (MAE), prediction error variance (PEV), and mean error (ME). The RMSE and MAE, and ME, respectively, are the square root of the sum of squared residuals, the sum of the absolute residuals, and the sum of residuals divided by the number of observations. The PEV is the variance of all the residuals. For a given observation from the data set, i.e., for a given section and at a given time period, the residual corresponds to the difference between the predicted and measured PSI.

The cumulative measures for each of the models are presented in Table 4. When looking at the results in absolute terms, it is important to remember that, even though the forecast horizon is only two years, the pavement sections were subjected to accelerated traffic loading. As expected, the IM model provided the smallest values for RMSE and MAE. RE and SUTSE, which consider heterogeneity, had the second and third smallest RMSE and MAE values. The benefits of considering heterogeneity are shown by the fact that RE performed better than OLS and SUTSE performed better than SE. The measures of the discrete model (OP) were similar to the OLS and SE models. The AASHO model and its updates had significantly larger RMSEs and MAEs. It was surprising that AASHO(T) and AASHO(D) performed worse than AASHO. A possible explanation is that the two models only update the load-applications-to-failure (ρ) but not the deterioration rate (β). Although AASHO(T) and AASHO(D) considered the censored data and predicted the load-applications-to-failure more accurately, the overall performance when the overall deterioration processes were compared was poor. If β is also re-estimated, AASHO(T) and AASHO(D) are expected to have improvement. Note that the ranking of MAE is very similar to that of RMSE, which means that the comparison is not greatly influenced by the outliers, i.e., it is robust.

Since $(RMSE)^2=PEV+(ME)^2$, analysis of PEV and ME^2 provides further insights. From the magnitudes of PEV and ME^2 in Table 4, it is clear that the ranking of RMSE is mainly determined by PEV. For example, the IM model slightly overestimated the pavement performance; however, according to the lowest PEV, the predictions of IM were robust and the overall performance in terms of RMSE was the best. On the other hand, OP had the lowest ME but it ranked only fifth due to its high PEV. We can conclude that the prediction errors of OP are deviated but they cancel each other out. Other empirical findings from

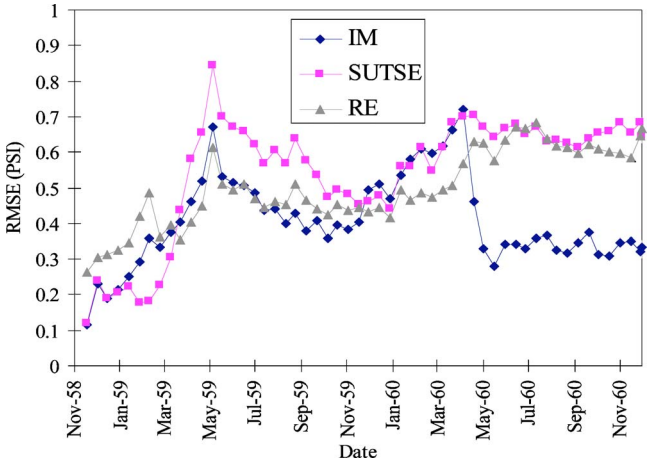

Fig. 1. RMSE comparison between IM, SUTSE, and RE

J. Infrastruct. Syst. 2008.14:138-149.

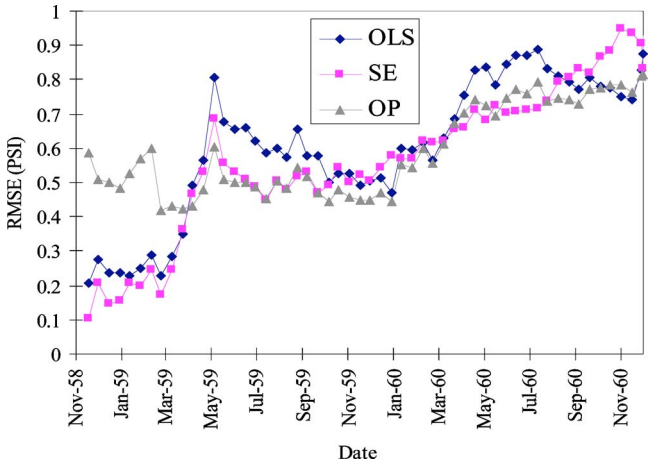

Fig. 2. RMSE comparison between OLS, SE, and OP

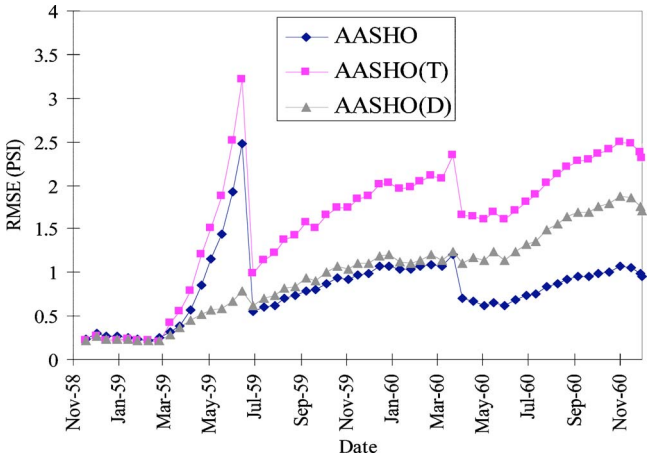

Fig. 3. RMSE comparison between AASHO, AASHO(T), and AASHO(D)

these two measures are the following. First, OLS and RE had very close PEV but OLS had higher ME, which means that the predictions of OLS were biased due to unobserved heterogeneity. Second, SUTSE provided lower PEV but higher absolute ME than SE. The lower PEV of SUTSE was expected due to the introduction of heterogeneity of variances. However, it is difficult to interpret the higher ME of SUTSE. The possible reasons are that the predictions of these models are influenced by many other factors such as serial dependence and latent performance and that ME has a much smaller impact on the overall performance.

Figs. 1–3 display the progression of RMSE over the duration of the test by mapping the accumulated ESAL for each section to the whole time horizon. In general, as the length of the forecast horizon increases, one expects the RMSEs to increase. It is important to note, however, that sections that fail (1) exhibit highly variable deterioration immediately prior to failure; and (2) are removed from the comparison data set after failure. As a result, it is difficult to anticipate what the RMSE trend should be. From Figs. 1–3, we observe that, as expected, the IM usually exhibited the smallest RMSE, and that it provided significantly better prediction in the later stages of the experiment or, equivalently, higher cumulative traffic for each section. Moreover, unlike any other models, the prediction capability did not seem to deteriorate for higher traffic. In Fig. 1, the RE and SUTSE models have similar patterns of RMSE. In Fig. 2, the OP model predicts worse than the OLS and SE models in the early stage. The three models perform almost identical afterward.

Fig. 3 shows an unexpected pattern of RMSE progression. For AASHO and AASHO(T), RMSE increased drastically after March 1959 until one of the sections was taken out of the test. Specifically, Section 299 (10.2-cm surface, 7.6-cm base, and 30.5-cm subbase layer), which provided very weak strength in relation to the traffic loads that were applied to it. AASHO predicted that the section would fail after 20 weeks; however, only one of the wheel paths failed and the section was taken out of the test after 34 weeks. Before the section was discarded, the AASHO model yielded negative PSI predictions with values as low as −25 PSI. The same pattern occured around March 1960 due to another pavement section. Clearly, the prediction of AASHO(D) was more robust than the other two. However, these models seemed to perform significantly worse than the other models in the present study.

Insights can be found by considering the discussions in the previous section and the empirical comparison in this subsection together. IM, RE, and SUTSE are superior to other models mostly because of the consideration of individual differences between pavement sections. This conclusion is particularly important because the data were collected in a highly controlled experiment. More interestingly, OLS is far superior to AASHO, AASHO(T),

Table 5. Aggregate-Level Comparison

Model	Section-specific correct (%)	Ranking	Very good (%)	Good (%)	Fair (%)	Poor (%)	Very poor (%)	Aggregate-level correct (%)	Ranking
AASHO	59.0	6	22.9	45.1	23.1	5.8	3.1	85.5	5
AASHO(T)	40.7	9	15.4	30.7	25.5	15.4	13.0	67.6	9
AASHO(D)	47.9	8	16.6	35.8	28.2	12.8	6.6	73.9	8
OLS	60.3	5	22.8	75.5	1.7	0.0	0.0	80.2	6
RE	61.1	4	7.2	79.9	12.8	0.1	0.0	79.7	7
IM	75.3	1	20.2	59.2	17.8	2.8	0.0	98.6	1
SUTSE	67.7	2	21.6	69.9	8.3	0.2	0.0	87.0	4
SE	62.1	3	12.8	55.5	30.2	1.5	0.0	87.5	3
OP	58.9	5	19.0	57.8	18.2	4.5	0.5	98.2	2
Observation	—	—	18.9	59.6	17.7	3.8	0.0	—	—

J. Infrastruct. Syst. 2008.14:138-149.

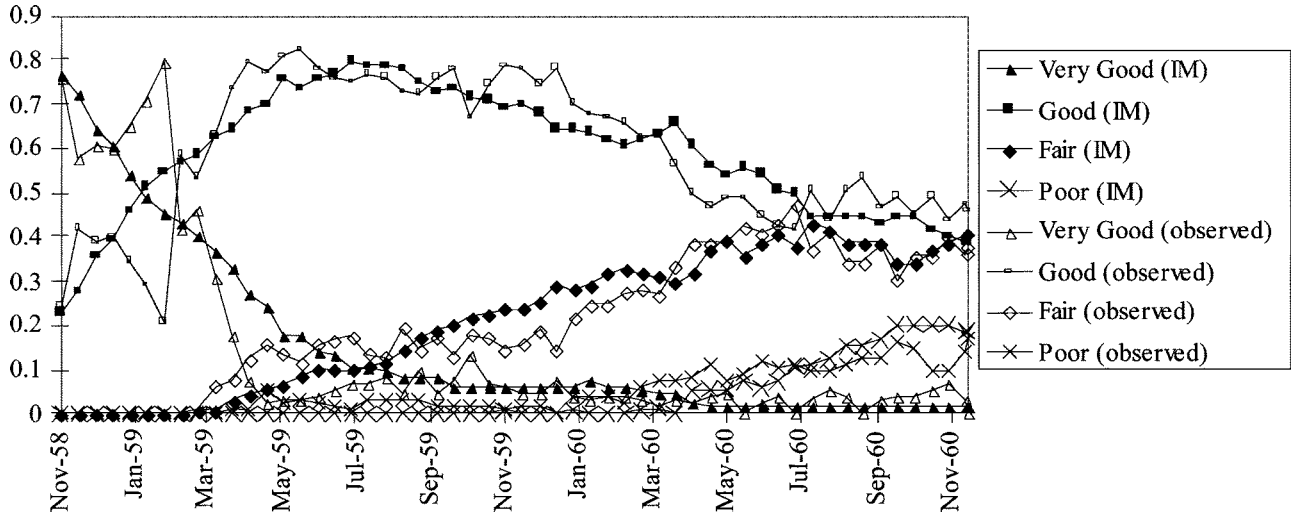

Fig. 4. Observed and predicted state probability (IM)

and AASHO(D) although they all adopt nonlinear regression formulation and use almost identical raw data. This constitutes additional evidence that the AASHO model was misspecified, and the fit-to-data was poor.

Aggregate/System-Level Comparison

The aggregate-level prediction and planning deal with systems of transportation facilities (AASHTO JTFP 2001). Two actions are usually taken at this managerial level. First, PSI (or condition) information for each facility may be far too detailed. Statistics such as PSI average or variance (over the facilities in the system) may be hard to interpret and not too informative. For this reason, facility condition is often described using elements from a finite set. Another reason is that when subjective manual inspection is used, accurate measurements are not available, and thus discrete ratings are often adopted. For example, the PSI can be used to assign discrete ratings as follows: very good ($4 \leqslant \text{PSI} < 5$), good ($3 \leqslant \text{PSI} < 4$), fair ($2 \leqslant \text{PSI} < 3$), poor ($1 \leqslant \text{PSI} < 2$), and very poor ($0 \leqslant \text{PSI} < 1$). Second, for the purpose of system-level decision making, the average percentages of facilities in each of the states can then be used as an aggregate/system-level performance measure and a so-called typical or average pavement is used to determine maintenance policies. In this subsection, we compare the models' capabilities to generate coarse, aggregate predictions, i.e., the predictions made by continuous models are converted into the aforementioned discrete ratings at both disaggregate and aggregate levels. We also compare the predictive capabilities of these models with those of the OP model, which is particularly well suited for situations where condition data are presented in the form of discrete categories.

The overall results of our study are presented in Table 5. In Table 5, the section-specific correct percent represents the percentage of correct prediction for individual sections, after the continuous PSI observations and predictions are converted into discrete ratings. The ranking of the section-specific correct percent is, in fact, similar to the ranking based on the continuous metric, RMSE. Therefore, for this example, the selection of statistical performance models can be made using continuous

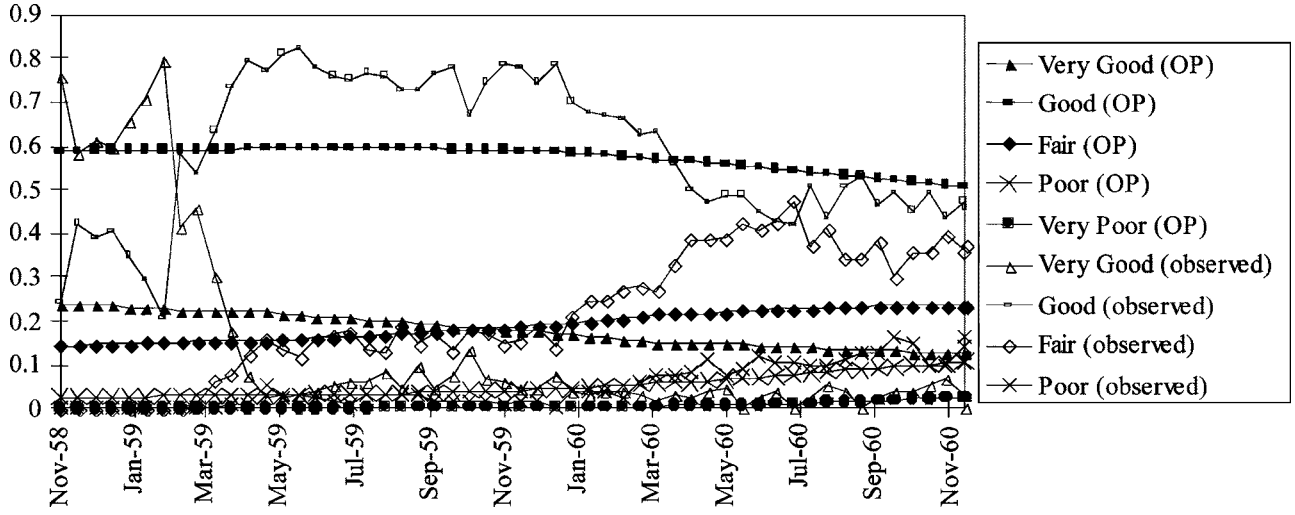

Fig. 5. Observed and predicted state probability (OP)

J. Infrastruct. Syst. 2008.14:138-149.

or discrete measures. However, whether 75.3% of correct prediction of IM, which is the highest possible percentage, is sufficient to make adequate maintenance policy is another topic for future research.

We further aggregate the section-specific predictions to the system level. The definitions of column headings are the following: very good, good, fair, poor, and very good means the average percentages of each state of all sections. Since the aggregate-level correct percent of each state is the smaller value among predicted percentage and observed percentage, the aggregate-level correct percent means the sum of the aggregate-level correct percent for all states. Once again, IM provides near-perfect predictions, which shows that it can also be used as an aggregate-level benchmark. Not surprisingly, the OP model is the second best model and the difference between the benchmark model is very small. We also observed that the predictions of SE, SUTSE, AASHO, and OLS are over 80% correct. RE, AASHO(D), and AASHO(T) performed the worst among the models. As shown in Table 5, the ranking of the aggregate correct prediction is substantially different from the ranking of the section-specific correct prediction and RMSE. Figs. 4 and 5 provide further interpretation of aggregate predictions. (IM predicts no pavement section in the very poor state for two years. Thus, the prediction of a very poor state is not shown in Fig. 4 to increase readability. Similarly, the observations for pavement sections in the very poor state are not shown in Figs. 4 and 5 because no pavement sections in the state were observed except one observation on March 23, 1960, and another one on April 20, 1960.) Figs. 4 and 5 track the change of predicted percentages over time for the IM and OP models. We can see that the IM predictions follow the observations, while the OP predictions attempt to capture the aggregate trends, and barely follow the actual observations. Therefore, we conclude that aggregate predictions can be misleading, and as a result, maintenance policies based on these predictions can be inadequate. This is consistent with the argument that an average or typical facility may not exist (Golabi and Shepard 1997).

Conclusions

We conducted an empirical comparison of the in-sample predictive capabilities of nine representative, state-of-the-art pavement performance models estimated with the functional performance data from the AASHO Road Test. The analysis provided a framework based on objective and quantitative measures to evaluate and select statistical performance models. Following are the conclusions and insights that stem from the results of the empirical study.

- The main observation that follows from the results of the disaggregate-level comparison is that the models that account for heterogeneity (differences between pavement sections in the panel) have significantly better predictive capabilities than the models that do not. The result is very significant because the models were estimated with data from a highly controlled experiment. We would expect the difference to be even greater when models are estimated with data sets that exhibit more variability, such as with data from facilities that are in service. While attractive for statistical and managerial reasons, model characteristics such as latent performance, incremental predictions, and serial dependence, do not display significant impact on the predictive capabilities of the models in the current study.
- AASHO(T) and AASHO(D) performed worse than AASHO because they only updated the load-applications-to-failure (ρ) but not the rate of deterioration (β). However, significant improvements by updating the AASHO model were not expected since the AASHO model was misspecified, which was proven by the fact that OLS was far superior to AASHO, AASHO(T), and AASHO(D) although they all adopt nonlinear regression formulation and use almost identical raw data in the estimation.
- The main observation that follows from the second part of the empirical study is that aggregate-level metrics can be misleading indicators of a model's predictive capabilities. We illustrated this point by comparing the rankings at disaggregate and aggregate levels and we found out that the two rankings differ substantially. Further, we contrasted the predictions of the IM and OP models over time and observed that the OP model attempts to capture the "aggregate" behavior across the facilities in the panel while IM is capable of tracking the actual observations very well. From a managerial perspective, this means that making maintenance and rehabilitation decisions based on aggregate-level measures could potentially lead to significant problems in the implementation of interventions because the "average" behavior may not be representative of the facilities comprising the system.
- The fact that the benchmark model, IM, only provided correct section-specific predictions for 75% of the observations and RMSE of 0.429 PSI is an indication that there is still room for developing models with improved capabilities. For example, the dynamic models used herein might be improved by adding additional autoregressive or moving average terms, or by including explanatory variables representing the raw data, as opposed to relying on the aggregate variables representing the pavement layer thicknesses or the traffic loading. Depending on the requirements, a different approach, such as mechanistic–empirical modeling, should be considered when the benchmark model does not provide satisfactory predictions.

Acknowledgments

This work was partially supported by the Northwestern University Transportation Center through a Dissertation Year Fellowship award to C.-Y.C. and by the National Science Foundation through Grant No. 0547471 awarded to P.L.D.-C. The work was done while C.-Y.C. was a graduate student at Northwestern University.

References

AASHTO. (1993). *AASHTO guide for design of pavement structures 1993*, AASHTO, Washington, D.C.

AASHTO Joint Task Force on Pavements (AASHTO JTFP). (2001). *Pavement management guide*, AASHTO, Washington, D.C.

Chu, C.-Y., and Durango-Cohen, P. L. (2007). "Estimation of dynamic performance models for transportation infrastructure using panel data." *Transp. Res., Part B: Methodol.*, 41(5), 493–505.

Frees, E. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*, Cambridge University Press, New York.

Gendreau, M., and Soriano, P. (1998). "Airport pavement management systems: An appraisal of existing methodologies." *Transp. Res., Part A: Policy Pract.*, 32(3), 197–214.

J. Infrastruct. Syst. 2008.14:138-149.

Golabi, K., and Shepard, R. (1997). "Pontis: A system for maintenance optimization and improvement of us bridge networks." *Interfaces*, 27(1), 71–88.

Golabi, K., Kulkarni, R. B., and Way, G. B. (1982). "A statewide pavement management system." *Interfaces*, 12(6), 5–21.

Highway Research Board. (1962). "The AASHO road test." *Special Rep. No. 61A-E*, National Academy of Science, National Research Council, Washington, D.C.

Li, Z. (2005). "A probabilistic and adaptive approach to modeling performance of pavement infrastructure. Ph.D. thesis, Univ. of Texas at Austin, Austin, Tex.

Madanat, S. M., Karlaftis, M. G., and McCarthy, P. S. (1997). "Probabilistic infrastructure deterioration models with panel data." *J. Infrastruct. Syst.*, 3(1), 4–9.

Madanat, S. M., Mishalani, R., and Ibrahim, W. H. W. (1995). "Estimation of infrastructure transition probabilities from condition rating data." *J. Infrastruct. Syst.*, 1(2), 120–125.

Madanat, S., Prozzi, J. A., and Han, M. (2002). "Effect of performance model accuracy on optimal pavement design." *Comput. Aided Civ. Infrastruct. Eng.*, 17, 22–30.

McNeil, S., Markow, M., Neumann, L., Ordway, J., and Uzarski, D. (1992). "Emerging issues in transportation facilities management." *J. Transp. Eng.*, 118(4), 477–495.

Mishalani, R. G., and Madanat, S. M. (2002). "Computation of infrastructure transition probabilities using stochastic duration models." *J. Infrastruct. Syst.*, 8(4), 139–141.

Paterson, W. D. O. (1987). *Road deterioration and maintenance effects: Models for planning and management*, Johns Hopkins University Press, Baltimore, MD.

Prozzi, J. A., and Madanat, S. M. (2000). "Analysis of experimental pavement failure data using stochastic duration models." *Transportation Research Record. 1699*, Transportation Research, Board, Washington, D.C., 87–94.

Prozzi, J. A., and Madanat, S. M. (2004). "Development of pavement performance models by combining experimental and field data." *J. Infrastruct. Syst.*, 10(1), 9–22.

Reinsel, G. (1984). "Estimation and prediction in multivariate random effects generalized linear model." *J. Am. Stat. Assoc.*, 79(386), 406–414.

Small, K. A., and Winston, C. (1988). "Optimal highway durability." *Am. Econ. Rev.*, 78(3), 560–569.

J. Infrastruct. Syst. 2008.14:138-149.