

Article

Review on Machine Learning Techniques for Developing Pavement Performance Prediction Models

Rita Justo-Silva ¹, Adelino Ferreira ^{1,*} and Gerardo Flintsch ²

¹ Research Centre for Territory, Transports and Environment, Department of Civil Engineering, University of Coimbra, 3030-788 Coimbra, Portugal; ritajustosilva@uc.pt

² Center for Sustainable Transportation Infrastructure, Virginia Tech Transportation Institute (VTTI), Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0131, USA; gflintsch@vtti.vt.edu

* Correspondence: adelino@dec.uc.pt

Abstract: Road transportation has always been inherent in developing societies, impacting between 10–20% of Gross Domestic Product (GDP). It is responsible for personal mobility (access to services, goods, and leisure), and that is why world economies rely upon the efficient and safe functioning of transportation facilities. Road maintenance is vital since the need for maintenance increases as road infrastructure ages and is based on sustainability, meaning that spending money now saves much more in the future. Furthermore, road maintenance plays a significant role in road safety. However, pavement management is a challenging task because available budgets are limited. Road agencies need to set programming plans for the short term and the long term to select and schedule maintenance and rehabilitation operations. Pavement performance prediction models (PPPMs) are a crucial element in pavement management systems (PMSs), providing the prediction of distresses and, therefore, allowing active and efficient management. This work aims to review the modeling techniques that are commonly used in the development of these models. The pavement deterioration process is stochastic by nature. It requires complex deterministic or probabilistic modeling techniques, which will be presented here, as well as the advantages and disadvantages of each of them. Finally, conclusions will be drawn, and some guidelines to support the development of PPPMs will be proposed.

Keywords: pavement performance prediction models; modeling techniques; machine learning



Citation: Justo-Silva, R.; Ferreira, A.; Flintsch, G. Review on Machine Learning Techniques for Developing Pavement Performance Prediction Models. *Sustainability* **2021**, *13*, 5248. <https://doi.org/10.3390/su13095248>

Academic Editor: Maria Rosaria De Blasiis

Received: 17 March 2021

Accepted: 3 May 2021

Published: 7 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Road transportation has always been inherent in the development of societies, impacting between 10–20% of Gross Domestic Product (GDP). Roads are responsible for personal mobility (access to services, goods, and leisure), and this is why world economies rely upon the efficient and safe functioning of transportation facilities.

Road maintenance is vital since the need for maintenance increases as road infrastructure ages. It is based on sustainability, meaning that spending money now saves much more in the future. Besides this, road maintenance plays a significant role in road safety. Nevertheless, since the budgets made available for maintenance are limited, pavement management is a challenging task. To select and schedule maintenance and rehabilitation operations, road agencies need to set up programming plans for the short term and the long term.

Although there is no doubt about the importance of PPPMs for efficient management in PMSs, these methods are not yet being used by most Portuguese road agencies or municipalities.

The purpose of this article is to present a review of past, present, and future modeling techniques used in the development of PPPMs. The assumptions, strengths, and weak points of each method and differences between them will be outlined. A brief introduction

to machine learning models will be presented, and the similarities and differences in statistics explained.

Finally, guidelines to help the development of PPPMs will be presented.

2. Pavement Performance Prediction Models (PPPMs)

A key component in pavement management is assessing the condition of the road network to predict future conditions.

Mathematical functions, known as pavement performance prediction models (PPPMs), are used to perform this task. PPPMs relate the pavement condition (e.g., cracking, rutting) to a set of explanatory variables (e.g., traffic loadings, age, environmental factors, pavement design characteristics). PPPMs can be classified (see Figure 1) according to [1]:

- Type of formulation (deterministic models, probabilistic models);
- Conceptual format (mechanistic, empirical, empirical–mechanistic);
- Application level (network level, project level); and
- Type of variables (dependent and independent).

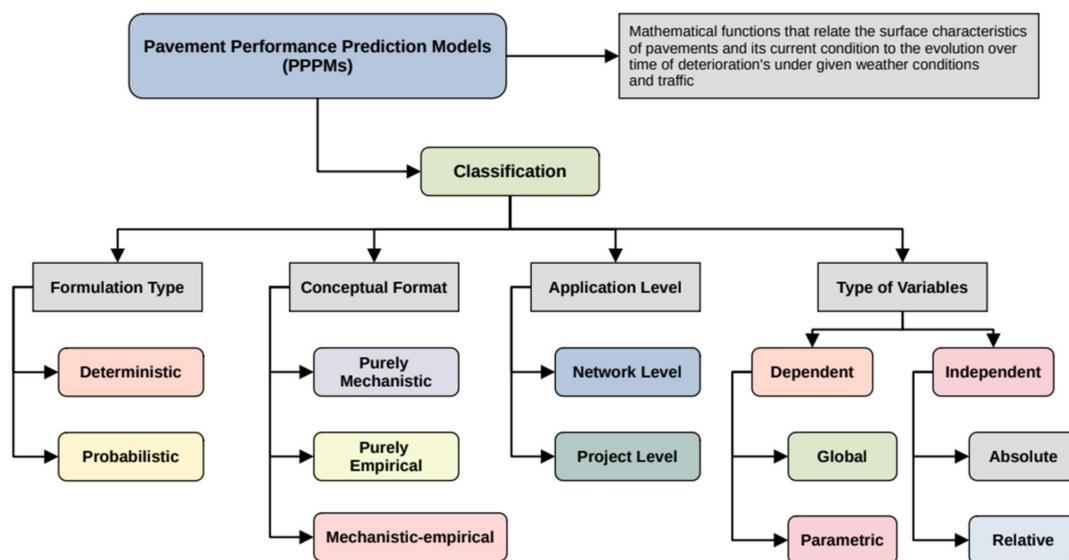


Figure 1. Flowchart of the classification of PPPMs.

At the project level, PPPMs are essential to evaluate the economic alternatives proposed (reconstruction, rehabilitation, and maintenance) to find the most cost-effective solution for each section. The level of detail and the amount of data is higher at the project level than at the network or strategic level. At the network level, PPPMs are used to predict the future class condition of the roads that comprise the network. According to the management level, some examples of the application of standard PPPM techniques are presented in Table 1.

According to [2], there are two types of models for pavement performance prediction:

- Static models (or absolute models); and
- Dynamic models (or relative models).

Table 1. Levels of application of PPPMs.

| Pavement Performance Prediction Models | | Management Level | | |
|--|----------------------------|------------------|-------------------|---------|
| | | National Network | Municipal Network | Project |
| Deterministic | Absolute | | | +++ |
| | Relative—structural | ++ | ++ | + |
| | Relative—functional | ++ | ++ | |
| Probabilistic | Bayesian methodology | ++ | ++ | |
| | Homogeneous Markov | +++ | ++ | |
| | Nonhomogeneous Markov | + | + | |
| | Semi-Markov process | + | + | |
| | Hidden Markov | + | + | + |
| Hybrid | Fuzzy logic | + | + | + |
| | Artificial neural networks | + | + | + |
| | Neuro-fuzzy | + | + | + |

Legend: +++ consistently used; ++ commonly used; + less used.

Static models do not take into account the lagged values of the output as inputs and can be described as

$$C_t = f(X_t, t) \quad (1)$$

where:

C_t = pavement condition at age t ; and

X_t = explanatory variables (e.g., structural characteristics, climatic conditions, traffic) at age t .

Typical examples of static models are regression models.

On the other hand, dynamic models forecast pavement performance using the lagged values of the pavement performance data and the explanatory variables, which should provide more accurate future predictions of pavement conditions. The use of dynamic models for developing PPPMs can be seen as modeling time series models and described as

$$C_t = f(C_{t-1}, \dots, C_{t-n}, X_t, X_{t-1}, \dots, X_{t-n}) \quad (2)$$

where:

C_t = pavement condition at age t ;

X_t = explanatory variables value at age t ; and

n = number of past observations considered.

Some authors claim that the stochastic nature of the pavement deterioration process, its nonlinear behavior, and the influence of unexplained explanatory variables require complex models to capture this deterioration process. Therefore, probabilistic models are prevalent in the United States of America and some European countries. Several countries also use deterministic models based on regression analysis and the Bayesian methodology.

The AASHTO road test occurred in the USA between 1958 and 1961, and it remains the foundation for the development of many PPPMs in various countries. It was the first major project carried out to analyze and predict the behavior of road pavements.

Several PPPMs were also developed globally, highlighting the models of HDM-4 [3], the SHRP Project, and the FORCE Project [4]. In Europe, the COST 324 Action [5] and the PARIS Project [6] remain essential references.

As mentioned before, PPPMs can be developed using different modeling techniques, depending on their formulation. The most common practices will be briefly presented in the following section.

3. Machine Learning Modeling Techniques for Developing PPPMs

Machine learning (ML) and statistics are intimately related fields in terms of methods but distinct in their principal goal: Statistics draw population inferences from a sample, whereas machine learning finds generalizable predictive patterns. ML algorithms use computational methods to “learn” information directly from historical data or experience. The algorithms adapt and improve their performance as the number of samples available for learning increases. ML algorithms have become popular nowadays since they can process and find natural patterns in large sets of data to make informed decisions based on better predictions. Machine learning models can be divided into three groups:

- Supervised learning—can be used for project-level or network-level pavement management;
- Unsupervised learning—can be used for exploratory and clustering analysis; and
- Reinforcement learning—can be used to help decision-makers for both project- and network-level pavement management.

3.1. Supervised Learning

Supervised learning uses input data and output data and builds a model to make useful predictions when applied to new data. If the goal is to predict a continuous output/target variable, like in project management, then regression machine learning techniques are used. The most common regression algorithms (see Table 2) include:

- Linear models;
- Nonlinear models;
- Decision trees (boosted and bagged);
- Neural networks; and
- Adaptive neuro-fuzzy learning.

Table 2. Summary of the most common SL regression algorithms (adapted from [7]).

| Regression Algorithm | How it Works | Best Used |
|--|---|--|
| Linear Regression | It is a statistical modeling technique used to describe a continuous response variable as a linear function of one or more predictor variables. Because linear regression models are simple to interpret and easy to train, they are often the first model to be fitted to a new data set. | When an algorithm that is easy to interpret and fast to fit is needed. As a baseline for evaluating other, more complex, regression models. |
| Nonlinear Regression | It is a statistical modeling technique that helps describe nonlinear relationships in experimental data. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. “Nonlinear” refers to a fitness function that is a nonlinear function of the parameters. | When data has strong nonlinear trends and cannot be easily transformed into a linear space. For fitting custom models to data. |
| Gaussian Process Regression Model | GPR models are nonparametric models that are used for predicting the value of a continuous response variable. They are widely used in the field of spatial analysis for interpolation in the presence of uncertainty. GPR is also referred to as Kriging. | For interpolating spatial data. As a surrogate model to facilitate optimization of complex designs such as automotive engines. |
| SVM Regression | Similar to SVM classification algorithms but are modified to be able to predict a continuous response. Instead of finding a hyperplane that separates data, SVM regression algorithms find a model that deviates from the measured data by a predefined value no greater than a small amount, with parameter values that are as small as possible (to minimize sensitivity to error). | For high-dimensional data (where there will be a large number of predictor variables). |

Table 2. Cont.

| Regression Algorithm | How it Works | Best Used |
|----------------------------------|---|---|
| Generalized Linear Models | It is a special case of nonlinear models that uses linear methods. It involves fitting a linear combination of the inputs to a nonlinear function (the link function) of the outputs. | When the response variables have nonnormal distributions, such as a response variable that is always expected to be positive. |
| Regression Trees | Similar to decision trees for classification, but they are modified to be able to predict continuous responses. | When predictors are categorical (discrete) or behave nonlinearly. |

If the data can be divided into groups or classes, and the goal is to predict a categorical/discrete output, classification machine learning is used. The most common algorithms for developing classification models (see Table 3) include:

- Logistic regression;
- Support vector machine;
- Decision trees (boosted and bagged);
- K-nearest neighbor;
- Neural networks;
- Naïve Bayes; and
- Discriminant analysis.

Table 3. Summary of the most common SL classification algorithms (adapted from [7]).

| Classification Algorithm | How it Works | Best Used |
|-------------------------------------|---|---|
| Logistic Regression | It fits a model that can predict the probability of a binary response belonging to one class or the other. Because of its simplicity, logistic regression is commonly used as a starting point for binary classification problems. | When data can be separated by a single, linear boundary. As a baseline for evaluating more complex classification methods. |
| K-Nearest Neighbor (kNN) | Categorizes objects based on the classes of their nearest neighbors in the data set. KNN predictions assume that objects near each other are similar. Distance metrics, such as Euclidean, city block, cosine, and Chebychev, are used to find the nearest neighbor. | When a simple algorithm to establish benchmark learning rules is required. When memory usage and prediction speed of the trained model is a lesser concern. |
| Support Vector Machine (SVM) | Classifies data by finding the linear decision boundary (hyperplane) that separates all data points of one class from those of the other class. The best hyperplane is the one with the largest margin between the two classes when the data is linearly separable. If the data is not linearly separable, a loss function is used to penalize points on the hyperplane's wrong side. SVMs sometimes use a kernel transform to transform nonlinearly separable data into higher dimensions where a linear decision boundary can be found. | For data with exactly two classes (can also be used for multiclass classification with a technique called error correcting output codes). For high dimensional, nonlinearly separable data. When a classifier that is simple, easy to interpret, and accurate is required. |
| Neural Networks | Inspired by the human brain, a neural network consists of highly connected networks of neurons that relate the inputs to the desired outputs. The network is trained by iteratively modifying the connections' strengths to map the given inputs to the correct response. | For modeling highly nonlinear systems. When data is available incrementally, and the goal is to update the model regularly. When model interpretability is not a key concern. |
| Naïve Bayes | Assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It classifies new data based on the highest probability of its belonging to a particular class. | For a small data set containing many parameters. When a classifier that is easy to interpret is needed. When the model encounters scenarios that were not in the training data, as is the case with many financial and medical applications. |

Table 3. Cont.

| Classification Algorithm | How it Works | Best Used |
|---|--|---|
| Discriminant Analysis | Classifies data by finding linear combinations of features. Discriminant analysis assumes that different classes generate data based on Gaussian distributions. Training a discriminant analysis model involves finding the parameters for a Gaussian distribution for each class. The distribution parameters are used to calculate boundaries, which can be linear or quadratic functions. These boundaries are used to determine the class of new data. | When a simple model that is easy to interpret is needed. When memory usage during training is a concern. When a model that is fast to predict is required. |
| Decision Tree | Predicts responses to data by following the tree's decisions from the root (beginning) down to a leaf node. A tree consists of branching conditions where the value of a predictor is compared to a trained weight. The number of branches and the values of weights are determined in the training process. Additional modification, or pruning, may be used to simplify the model. | When an algorithm that is easy to interpret and fast to fit is a requirement. To minimize memory usage. When high predictive accuracy is not a requirement. |
| Ensemble Methods (Bagged and Boosted Decision Trees) | Several "weaker" decision trees are combined into a "stronger" ensemble. A bagged decision tree consists of trees trained independently on data that is bootstrapped from the input data. Boosting involves creating a strong learner by iteratively adding "weak" learners and adjusting each weak learner's weight to focus on misclassified examples. | When predictors are categorical (discrete) or behave nonlinearly. When the time needed to train a model is less of a concern. |

3.2. Unsupervised Learning

Unsupervised learning is used to find patterns in data and draw inferences from data sets that only have input data. Unsupervised learning is often used for exploratory data analysis and clustering. The standard algorithms for unsupervised learning include:

- Gaussian mixture models;
- K-means and k-medoids;
- Hierarchical clustering;
- Hidden Markov models; and
- Fuzzy c-means.

Clustering works by grouping similar points using a distance metric. Even if the goal is to perform supervised learning, clustering can be an excellent tool for hypothesis development, modeling over smaller subsets of data, data reduction, and outlier detection. Clustering algorithms fall into two broad groups:

- Hard clustering—where each data point belongs to only one cluster (see Table 4); and
- Soft clustering—where each data point can belong to more than one cluster (see Table 5).

Hard or soft clustering techniques can be used if possible data groupings are known.

Table 4. Summary of the most common UL hard clustering algorithms (adapted from [7]).

| Classification Algorithm | How it Works | Best Used |
|--------------------------------|--|--|
| K-Means | Partitions data into k number of mutually exclusive clusters. How well a point fits into a cluster is determined by the distance from that point to the cluster's center. RESULT = cluster centers. | When the number of clusters is known. For fast clustering of large data sets. |
| K-Medoids | Similar to k-means, but with the requirement that the cluster centers coincide with points in the data. RESULT = cluster centers that coincide with data points. | When the number of clusters is known. For fast clustering of categorical data. To scale to large data sets. |
| Hierarchical Clustering | Produces nested clusters by analyzing similarities between pairs of points and grouping objects into a binary, hierarchical tree. RESULT = dendrogram showing the hierarchical relationship between clusters. | When the number of clusters in data advances, a visualization to guide selection is desirable. |
| Self-Organizing Map | Neural network-based clustering that transforms a data set into a topology-preserving 2D map. RESULT = lower-dimensional (typically 2D) representation. | To visualize high-dimensional data in 2D or 3D. To deduce the dimensionality of data by preserving its topology (shape). |

Table 5. Summary of the most common UL soft clustering algorithms (adapted from [7]).

| Classification Algorithm | How it Works | Best Used |
|-------------------------------|--|--|
| Fuzzy c-Means | Partition-based clustering when data points may belong to more than one cluster. RESULT = cluster centers (similar to k-means) but with fuzziness so that points may belong to more than one cluster. | When the number of clusters is known. For pattern recognition. When clusters overlap. |
| Gaussian Mixture Model | Partition-based clustering where data points come from different multivariate normal distributions with specific probabilities. RESULT = a model of Gaussian distributions that give probabilities of a point being in a cluster. | When a data point might belong to more than one cluster. When clusters have different sizes and correlation structures within them. |

3.3. Reinforcement Learning

Reinforcement learning, unlike supervised and unsupervised learning, works with data from a dynamic environment. The goal is to find the best sequence of actions that will produce the most reward in the long run. The agent/algorithm explores, interacts with (through actions), and learns from the environment to determine the best policy. Reinforcement learning can be divided into two main groups:

- Model-based reinforcement learning; and
- Model-free reinforcement learning.

In Figure 2, the summary of ML algorithms is presented.

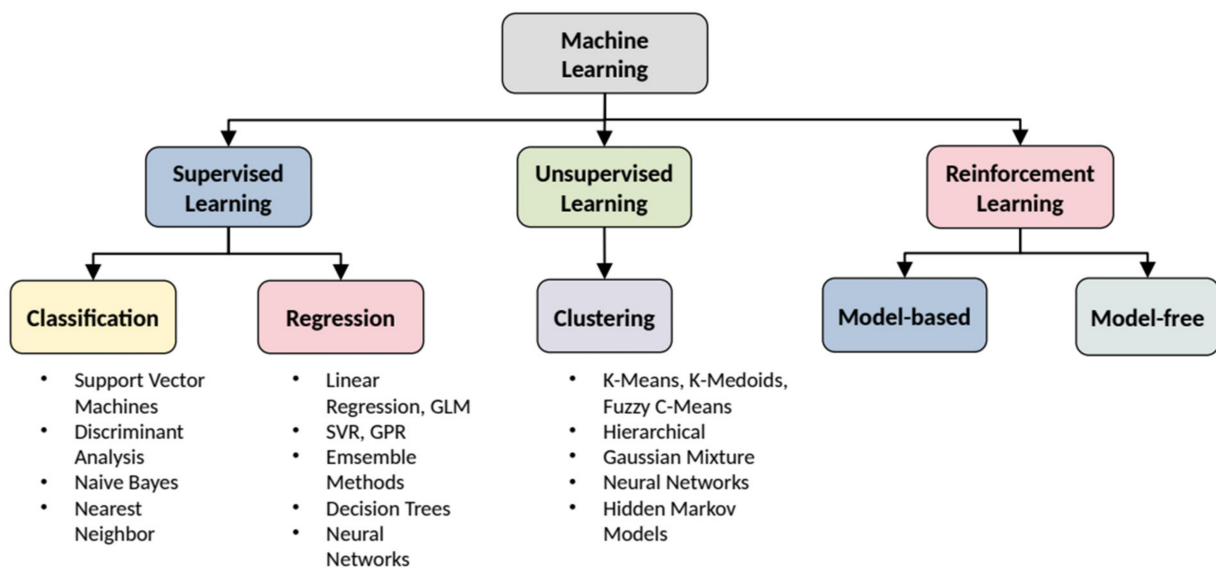


Figure 2. Summary of machine learning algorithms (adapted from [7]).

Machine learning algorithms can also be divided [8] according to:

- Information-based learning;
- Similarity-based learning;
- Probability-based learning;
- Error-based learning.

4. Data Pre-Analysis, Visualization, and Preparation

4.1. Data Pre-Analysis

Before attempting to build predictive models, it is important to understand the type of data under analysis. The main data types are:

- Numerical data—represents data/information that is measurable, which can be divided into two subcategories:
 - Discrete—integer-based data (e.g., M&R actions, number of pavement sections); and
 - Continuous—decimal-based data (e.g., pavement structural capacity, traffic, pavement condition);
- Categorical data—qualitative data that are used to classify data by categories (e.g., crack initiation = true or false); and
- Ordinal data—represent discrete and ordered data/information (e.g., rank position = 1st, 2nd, 3rd; rutting level = low, medium, high).

After understanding the type of data involved, the next step is to make an exploratory analysis and, if necessary, perform some data preparation. There are two goals in data exploration:

1. To fully understand the characteristics of each variable in data (types of values the variable can take, the ranges into which the values fall, and how the values are distributed across that range); and
2. To discover any data quality issues (which may arise due to invalid data or perfectly valid data that may cause difficulty to some machine learning techniques).

The most common data quality issues are:

- Missing values—if features have missing values, it is necessary to understand why they are missing. For example, road agencies usually do not make pavement inspections every year, rather every two, three, or four years;

- Irregular cardinality problems—continuous features will usually have a cardinality value close to the number of instances in the data set. If the cardinality of a continuous feature is significantly less than the number of instances in the data set, it should be investigated; and
- Outliers—values that lie far away from the central tendency and can represent valid or invalid data. Valid outliers are correct values that are simply very different from the rest of the values for a feature and should not be removed from the analysis. In contrast, invalid outliers are often the result of noise in the data (sample errors) and must be removed.

Some machine learning techniques do not perform well in the presence of outliers; consequently, it is essential to identify outliers and know how to deal with them.

Developing a data quality report is the most crucial tool of the data exploration process. It should include the characteristics of each feature using:

- Standard measures of central tendency (mean, mode, and median);
- Standard measures of variation (standard deviation and percentiles);
- Standard data visualization plots (bar plots, histograms, and box plots).

4.2. Data Visualization

The histograms of features allow us to relate shapes of well-understood probability distributions (see Figure 3), which help to define ML models.

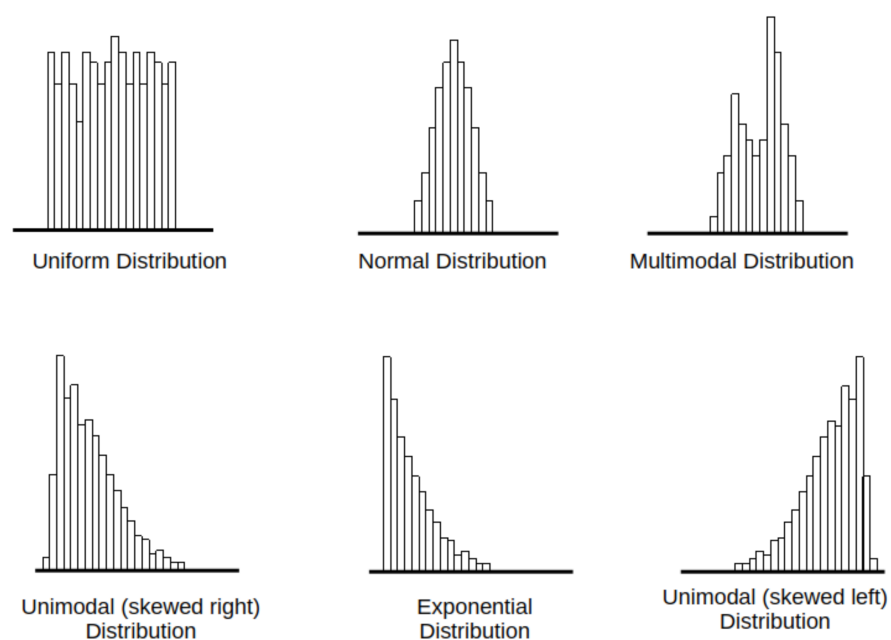


Figure 3. Histograms of features of some probability distributions (adapted from [8]).

A uniform distribution indicates that a feature is equally likely to take any value within its range.

Features following a normal distribution (unimodal) are characterized by a strong tendency toward a central value and symmetrical variation to either side of this central tendency. Finding features that exhibit a normal distribution is advantageous since many modeling techniques work particularly well with normally distributed data.

A feature characterized by a multimodal distribution has two or more very commonly occurring ranges of values that are separated. Multimodal distributions tend to occur when a feature contains a measurement made across a few distinct groups.

Unimodal histograms that exhibit skew illustrate a tendency toward very high or very low values.

Finally, in a feature following an exponential distribution, the likelihood of low values occurring is very high but diminishes rapidly for higher values. Exponential distributions are likely to contain outliers.

4.3. Data Preparation

Data preparation allows us to change how data is represented to make it more suitable for ML algorithms. The three most common techniques for data preparation are:

- Normalization (range normalization, standard scores)—aims to prepare descriptive features to fall in particular ranges;
- Binning (equal width, equal frequency)—involves converting continuous features into categorical features; and
- Sampling (top, random, stratified)—consists of taking a representative data sample from the original (larger) data set.

The development of the PPPMs is the next step. Therefore, the modeling techniques are described in the following section. In Figure 4, a workflow of the development of PPPMs is presented.

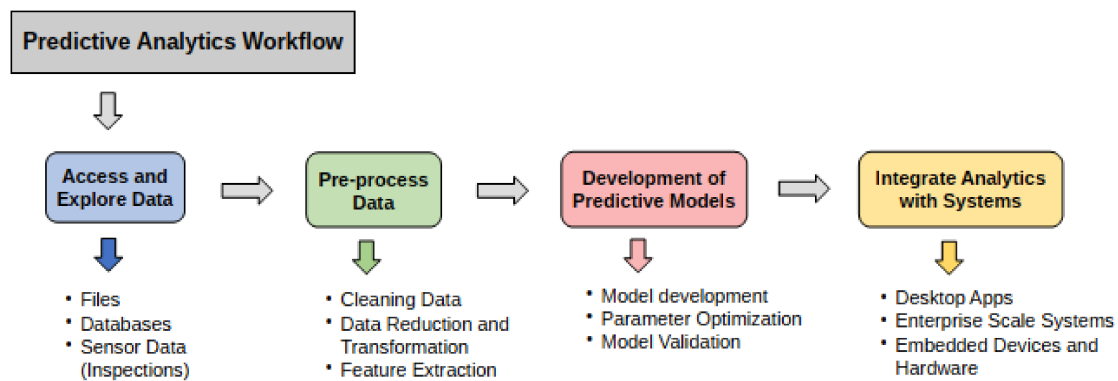


Figure 4. Predictive analytics workflow (adapted from [7]).

5. Information-Based Models

Information-based models aim to determine the input features that provide the most information about the target feature (dependent variable). In terms of PPPMs, several input features such as type of pavement, structural capacity, traffic, or age can be used to predict the pavement condition.

Claude Shannon's model of entropy is used to measure the information gain of the input features. Decision trees and model ensembles are examples of information-based models. Decision tree models can model the interactions between explanatory features and can be used for data sets that contain both categorical and continuous input variables. However, decision trees tend to become quite large when the input variables are continuous, decreasing the model's interpretability. Model ensembles generate a group of models and then make the predictions by aggregating the output from those models.

The two standard approaches for model ensembles are boosting and bagging. More detailed information on information-based models can be found in [8].

6. Similarity-Based Models

Similarity-based models use measures of similarity and feature spaces to make predictions.

The two most commonly used distance metrics are the Euclidean (see Equation (3)) and the Manhattan distance (see Equation (4)), which are particular cases of the Minkowski distance.

$$\text{Euclidean}(a, b) = \sqrt{\sum_{i=1}^m (a_i - b_i)^2} \quad (3)$$

$$\text{Manhattan}(a, b) = \sum_{i=1}^m |a_i - b_i| \quad (4)$$

The nearest-neighbor algorithm is an example of a similarity-based model.

7. Error-Based Models

In error-based machine learning, the goal is to search for a set of parameters for a parameterized model that minimizes the total error across the model's predictions using a set of training instances.

7.1. Linear Regression Models

Linear regression is a statistical tool that analyzes the relationship between a single dependent variable (criterion) with one or a set of independent variables (predictors or explanatory variables). The most well-known and straightforward mathematical model that can capture the relationship between two continuous features is the line equation.

When more than one explanatory variable exists, the simple regression is called multiple linear regression. The main objective is to predict values of the dependent/target variable under study (e.g., cracking, rutting), knowing the explanatory variables (e.g., traffic, age, structural capacity). Each explanatory variable is weighted by the regression procedure to ensure the model's maximum prediction and denote the relative contribution of each one. In a multiple linear regression model, the relationship between the dependent variable and the various independent variables is assumed to be linear and defined by Equation (5):

$$y(x, w) = w_0 + w_1 x_1 + \dots + w_n x_n + \varepsilon = \sum_{i=0}^n w_i x_i + \varepsilon \quad (5)$$

where:

$y(x, w)$ = value of the predicted target/output variable;

x_i = values of the explanatory/input variables;

w_0 = intercept, which represents the value of the target variable when x_i is 0;

w_i = regression coefficients (represent the extent to which the input variables are associated with the target variable); and

ε = disturbance term (represents the random error associated with the regression).

The key to developing linear regression models is to determine the optimal values for the weights in the model, which allow the model to best capture the relationship between the explanatory features and a target feature.

It is important to note that the model is built using a sample, which is used to make inferences about the total population data.

The coefficients' values are determined by minimizing an error function that measures the misfit between the predicted output/target values obtained by the model $y(x, w)$ and the observed target values y in the data set. There are several error functions, but the most commonly used is the sum of squares of the errors, defined by Equation (6):

$$E(w) = \frac{1}{2} \sum_{i=1}^n [y(x, w) - y]^2 \quad (6)$$

The least square error approach for finding the model parameters w represents a specific case of maximum likelihood.

Each pair of weights $w[0]$ and $w[1]$ defines a point on the $x - y$ plane, and the sum of squared errors for the model, using these weights, determines the height of the error surface above the $x - y$ plane for that pair of weights. The $x - y$ plane is known as the weight space, and the surface is the error surface.

The model that best fits the training data is the model corresponding to the lowest point on the error surface, i.e., the global minimum, which corresponds to the point at which the partial derivatives of the error surface (concerning the weights) are equal to zero.

However, as the number of explanatory variables and consequently the number of weights increases, the brute-force search approach to finding the optimal set of weights becomes unfeasible. A clever way is the use of algorithms such as the gradient descent algorithm to perform this task by:

- Starting with a set of random weight values;
- Iteratively making small adjustments to these weights based on the output of the error function. It is supposed that the errors show that the predictions made by the model are higher than the observed values. In this case, the weight should be decreased if the explanatory variable positively impacts the target variable; and
- According to the gradient of the error surface, the algorithm moves downwards on the error surface at each step (using differentiation and partial derivatives) to converge.

The values chosen for the learning rate and initial weights can significantly impact how the gradient descent algorithm proceeds. Unfortunately, there are no theoretical results that help in selecting the optimal values for these parameters. Instead, these algorithm parameters must be chosen using rules of thumb gathered through experience. The learning rate α in the gradient descent algorithm determines the size of the adjustment made to each weight at each step in the process.

In linear regression models, the associated error surfaces are determined by the model's linearity rather than the data properties. Therefore, linear regression models present two fundamental properties that allow us to find the optimal combination of weights:

- They are convex (the error surfaces are shaped like a bowl); and
- They have a global minimum (meaning a unique set of optimal weights with the lowest sum of squared errors on an error surface).

The advantages of using regression analysis are:

- It is suitable for modeling a wide variety of relationships between variables;
- In many practical applications, the assumptions of linear regression are often suitably satisfied;
- Its outputs are relatively easy to interpret and communicate; and
- The estimation of regression models is relatively easy. The routines for its computation are available in a vast number of software packages.

However, regression analysis follows some assumptions that must be verified, such as:

- The continuous behavior of the target variable;
- The linearity relationship between the target and explanatory variables;
- The behavior of disturbance terms (not auto-correlated, no correlation with the regressors, and the normally distributed pattern).

Regression analysis is probably the most widely used method for pavement performance prediction. The AASHTO pavement design equations [9] are an excellent example of using regression analysis for pavement performance predictions. Additional examples of the application of regression models can also be found in [10].

Even though regression models are based on a large body of research and best practice in statistics, they can be extended in many ways.

7.2. Logistic Regression Models

The linear regression models described previously assume that the target/dependent variable is continuous. However, in pavement management, it is useful, for example, to

assess whether a road pavement has deteriorated beyond a particular threshold, which sets the target variable as a binary outcome (1 or 0).

$$Y_i = \text{logit}(P_i) = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_K X_{K,i} \quad (7)$$

where:

Y_i = value of the predicted target/output variable;
 X_k = set of the explanatory/input variables;
 β_0 = model constant; and
 β_K = set of unknown parameters.

Logistic regression models allow the prediction of categorical targets rather than continuous ones by placing a threshold on the multiple linear regression model's output variable, using the logistic function presented in Equation (8). An alternative to logistic regression is probit regression, which relies on normal distribution instead of logistic distribution.

$$M_w(x) = \text{logistic}(wx) = \frac{1}{1 + e^{-wx}} \quad (8)$$

The logistic regression model is logarithmic at extreme values and approximately linear in the middle ranges (S-shaped curve). The output of logistic regression models can be interpreted as the probability of the presence of a particular class of pavement condition level.

$$P(Y_i|C_n) = \frac{e^{\beta_c X_i}}{\sum_{k=1}^K e^{\beta_k X_i}} \quad (9)$$

Models can be set as:

- Ordinal (if the order of the target variable is important)—the final model predicts the same regression coefficients but different intercepts for each class;
- Nominal (if the order of the target variable is not essential)—the final model predicts different regression coefficients and intercepts for each class.

The multinomial logit model (MNL) is an extension of the logistic regression models for more than two alternatives. For k target levels (pavement condition classes), $k - 1$ different logistic regression models are built since one of the pavement condition classes is chosen as the reference category.

The unknown parameters in each vector β_k are estimated iteratively by the maximum a posteriori (MAP) estimation, which is an extension of the maximum likelihood using the regularization of the weights.

The MNL assumes that each independent variable has a single value for each case. It also assumes that the dependent variable cannot be perfectly predicted from the independent variables. Besides, there is no need for the independent variables to be statistically independent of each other (unlike, for example, in a naïve Bayes classifier). However, collinearity is assumed to be relatively low.

Several logistic regression models have been applied to pavement management to predict pavement conditions [11–14].

7.3. Nonlinear Regression Models

The simple linear regression and logistic regression models only represent linear relationships between descriptive features and a target feature. In many cases, this assumption limits the creation of an accurate prediction model.

By applying a set of basis functions to descriptive features, models representing nonlinear relationships can be created. The advantage of using basis functions is that they allow models representing nonlinear relationships to be built even though these models remain a linear combination of inputs. Consequently, it is still possible to use the gradient descent algorithm to train them. The main disadvantages of using basis functions are:

- The set of basic functions must be manually inputted; and
- The number of weights in a model using basis functions is usually far greater than the number of descriptive features. Therefore, finding the optimal set of weights involves searching through a much broader set of possibilities (i.e., a much larger weight space).

To assess the need to use a nonlinear model, a plot of the target/output variable to each input/explanatory variable can be done. Before building a nonlinear model, it is also useful to transform the input and output variables such that the relationship between the transformed variables is linear. Nonlinear models such as nonlinear ARX or Hammerstein–Wiener models can be developed if the variable transformations that yield a linear relationship between input and output variables cannot be found. However, a linear model is often suitable for describing the system dynamics accurately, and, in most cases, it should be the starting point before developing more complex models.

7.4. Time-Series Models

A time-series is a sequence of observations arranged by their time of outcome. Time-series models have been the focus of considerable research [15–19] and development in recent years. This interest results from the insights gained when observing and analyzing the behavior of variables over time, allowing future outcomes to be forecast.

A fundamental property that sets time-series methods apart from other approaches is that time-series data are not independently generated. Hence, procedures that assume independently and identically distributed data are unsuitable.

When analyzing time-series data, time-domain or frequency-domain approaches are often used. The time-domain approach assumes that adjacent points in time are correlated and that future values are related to past and present ones.

The frequency-domain approach assumes that time-series characteristics relate to periodic or sinusoidal variations reflected in the data.

Additionally, time-series analysis techniques may be divided into parametric and nonparametric methods. The parametric approaches assume that the underlying stationary stochastic process has a specific structure, which can be described by a small number of parameters (for example, using an autoregressive or moving average model). In these approaches, the goal is to estimate the parameters of the model, which describe the stochastic process. By contrast, nonparametric techniques estimate the covariance without assuming that the process has any particular structure. Methods of time-series analysis may also be divided into:

- Linear/non-linear; and
- Univariate/multivariate.

A time series is one type of panel data. Panel data are a general class, a multidimensional data set, whereas a time-series data set is a one-dimensional panel (as is a cross-sectional data set). A data set may exhibit characteristics of both panel data and time-series data. One way to differentiate between the two is to determine what makes one data observation unique from the other observations. If the answer is the time data variable, this is a time-series data set. If determining a unique observation requires a time data variable and an additional identifier unrelated to time (section ID, section location), it is panel data. If the differentiation lies in the non-time identifier, the data set is a cross-sectional data set.

7.5. Panel/Longitudinal Data Models

Traditionally, statistical and econometric models have been estimated using cross-sectional or time-series data. A typical cross-section represents several data sections concerning a particular year, whereas time series show different time periods for one section. However, if data are available based on cross-sections of individuals observed over time, these data, which combine cross-sectional and time-series characteristics, are called panel data.

Panel data models, also named longitudinal data models, allow researchers to construct and test realistic behavioral models that cannot be identified using only cross-sectional or time-series data.

In longitudinal data road studies, the hypothesis that observations are independent and identically distributed is no longer valid. Therefore, it is necessary to use mixed models, which assume two sources of variation within and between sections.

Panel data models are widely used for repeated measurements. Mixed-effects models offer a flexible framework where the population characteristics are modeled as fixed effects, and unit-specific variation is modeled as random effects [20].

The fixed effects model investigates the relationship between the predictor and the outcome variables within an entity, which have characteristics that may or may not influence the predictor or outcome variables. Each entity is different. Therefore, the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others. Moreover, the variable-intercept models consider entities or time (one-way models) or both entities and time (two-way models). Fixed effects are the simplest and most straightforward models for accounting for cross-sectional heterogeneity in longitudinal data.

The random effects model states that the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model. When the differences across entities have some influence on the dependent variable, random effects should be used.

A panel data regression differs from a regular time series or cross-section regression as it has a double subscript on its variables, as shown in Equation (10).

$$Y_{it} = \alpha + X_{it}\beta + u_{it} \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (10)$$

where:

i refers to the cross-sectional units;

t refers to the time periods;

α is a scalar;

β is a vector;

X_{it} is the i^{th} observation on K the explanatory variable;

u_{it} is the error component.

The advantages of panel data models are:

- Controlling for individual heterogeneity: Panel data suggest that entities are heterogeneous, whereas studies of time-series and cross-section do not control this heterogeneity, which may lead to biased results;
- More informative data: Panel data give more variability, less collinearity among the variables, more degrees of freedom, and more efficiency;
- The ability to study the dynamics of adjustment: Panel data are better for this. Cross-sectional distributions that look relatively stable hide a multitude of changes; and
- Identify and measure the effect that is simply not detectable in pure cross-sectional or pure time-series data, allowing more complex behavioral models to be constructed and tested than with pure cross-sectional or time-series data.

The use of panel data models also includes some limitations such as heterogeneity, correlation in the disturbance terms, and heteroscedasticity. These disadvantages must be accounted for during the analysis. They could be related to groups with similar behavior among their elements and with significantly different behavior from other groups.

Linear mixed-effects models (LME) were successfully used by [21–23] as well as non-linear mixed effects models (NLME) by [24,25].

Archilla and Madanat [26] proposed a linear mixed-effects model for road pavements.

Lorino et al. [20] developed a nonlinear mixed-effects model for describing pavement section behavior as a function of time, taking into account a logistic function. The aim was

to model the sigmoid evolution law of pavement cracking and incorporate one covariate into the model to examine the climate factor's effects on pavement behavior.

7.6. Support Vector Machines

Vladimir Vapnik and Alexey Ya Chervonenkis invented support vector machines (SVM) in 1963 to address a problem related to logistic regression models.

Logistic regression attempts to maximize the probability of the classes of known data points according to the model. Therefore, the classification boundary arbitrarily may be placed close to a particular data point, which disregards the common-sense notion that a good classifier should not set a boundary near a known data point (data points that are close to each other should be part of the same class). On the other hand, support vector machines are non-probabilistic, so they assign a data point to a class with 100% certainty.

Support vector machines work by constructing a hyperplane that separates points between two classes. The hyperplane is determined using the maximal margin hyperplane, which is the hyperplane that represents the maximum distance from the training observations. SVMs can be defined as linear classifiers with the following two assumptions:

1. The margin should be as wide as possible; and
2. The support vectors (data points from each class that lie closest to the classification boundary) are the most useful data points because they are most likely to be incorrectly classified.

The second assumption of SVMs is fundamental since this means that after the training phase, the SVM only performs classification using the support vectors instead of considering the entire data set.

Another essential property of SVMs is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum.

Karballaezadeh et al. [27] developed an SVM model to estimate the remaining service life of a pavement.

Ziari et al. [28] in their research analyzed five kernel types of SVM algorithms to predict the future of the pavement condition using the international roughness index (IRI) as the pavement performance index.

7.7. Artificial Neural Networks

Artificial neural networks (ANNs) are computational systems inspired by biological and psychological insight composed of processing elements, called "neurons." Neurons are linked to each other, establishing a network. The strength of the connection between neurons is called "weight." To process information, neurons take several inputs, weigh them, sum them up, and then give a weighted sum of the inputs to the network as output. In ANNs, neurons are usually organized in "layers." Layers consist of weights and the subsequent neurons that sum up the signals they carry [29]. A typical ANN (see Figure 5) has an input vector, one or more hidden layers, and an output layer. Information flows from the input vector to the hidden layers and from the hidden layers to the output layer. This technique can learn with the data and, when it is well trained, can estimate the results based on the inputs without understanding the relations between them, and does not require algorithms or experts in the field [30]. Training is accomplished by sequentially applying input vectors while adjusting network weights according to a predetermined procedure until we have a consistent output set [29].

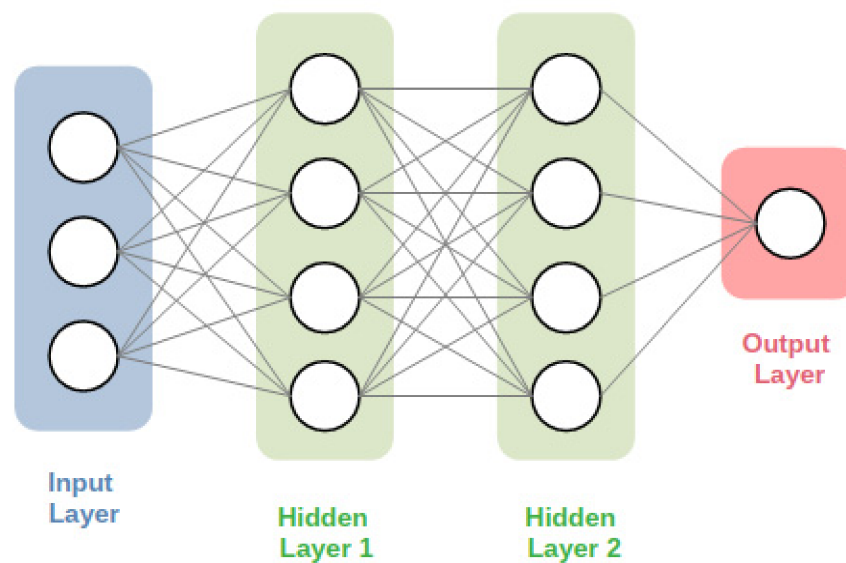


Figure 5. Structure of a neural network.

ANNs are used in the Transportation Department of Arizona to manage conservation actions [31] and in the Transportation Department of Kansas to predict the roughness of the pavement [30]. ANNs are suited to solving complex problems and can adapt to dynamic environments in real time. Even so, ANNs are data-driven systems, and if the training process is not done correctly, the network may suffer from an incomplete representation of the data or over-training [32].

ANNs are an excellent tool for dealing with the complexity of pavement structures and the inherent non-linearity of the measured data.

Expressing a complex system through this powerful technique has proven to successfully overcome many of the limitations of classical methods such as finite elements and traditional statistical analyses [33,34]. Nevertheless, these systems cannot explain decisions since they are obtained by a simultaneous execution of a large number of neurons, which is usually very hard to interpret [35].

8. Probability-Based Models

Probability-based models are based on the probability theory and Bayes' theorem. The most common techniques are described in the following sections.

8.1. Naïve Bayes Model

Naïve Bayes' methods are supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features and model the features as conditionally independent of the class. Consequently, naïve Bayes classifiers are highly scalable and can quickly learn to use high-dimensional features with limited training data. This feature is helpful for many real-world data sets where the amount of data is small compared to the number of features for each piece of data, such as speech, text, and image data.

8.2. Bayesian Networks

Bayesian methodology allows a combination of objective data (obtained from visual inspections) and subjective data (opinion of experts in this area) to develop PPPMs. This approach can also be used to create equations exclusively from subjective information.

Bayesian networks use a graph-based representation to encode the structural relationships, such as direct influence and conditional independence between subsets of features in a domain. Consequently, a Bayesian network representation is more compact than a full joint distribution (because it can encode conditional independent relationships), yet it is

not forced to assert global conditional independence between all descriptive features [8]. Bayesian network models are an intermediary between full joint distributions and naïve Bayes models and offer a practical compromise between model compactness and predictive accuracy.

The use of Bayesian methodology is not new. Smith et al. [36] developed a model relating pavement distresses with various design variables.

Haper and Majidzadeh [37] exploited this technique to include information elicited from experts on PMSs.

In their research, Hajek and Bradbury [38] used this methodology to develop a PPPM for asphalt concrete surfaces containing steel slag aggregates.

Bayesian methodology was also used in the PMS of JAE (Portuguese Road Administration) by Pereira and Barbosa [39] to include information based on experts' knowledge in calculating transition probabilities, which were updated when new data from pavement inspections were available.

Hong and Prozzi [40] analyzed and updated an existing incremental pavement deterioration model based on data from the AASHTO Road Test and presented the Bayesian approach to estimate the model, using the Gibbs sampling algorithm and Monte Carlo Markov chain simulation to estimate the distribution of each parameter.

More recently, Jiménez and Mrawira [41] used a Bayesian regression to predict rut depth progression for pavement deterioration modeling based on the AASHTO Road Test.

In his dissertation, Liao [42] devised a novel approach for developing performance prediction models for pavements that received preservation treatments. The data for developing and testing this model was obtained from the long-term pavement performance (LTPP) database. Artificial neural networks (ANNs) and Bayesian regression techniques were employed to develop the components of this model.

8.3. Markov Models

Markov models are a worldwide reference for describing the pavement deterioration processes.

8.3.1. The Homogeneous Markov Process

The pavement deterioration process is known to be stochastic because of measurement errors, non-linearity behavior in the deterioration process, and the influence of unexplained explanatory variables. It requires complex deterministic models to capture the deterioration process. Therefore, according to some authors, pavement performance prediction should be based on probabilistic models rather than deterministic ones [43–47]. Using the empirical stochastic-based approach in the design of flexible pavement is also justifiable [48].

The Markov process is a stochastic description of event development that is assumed to be time independent. The process of the deterioration of pavements is given by a matrix of transition probabilities [45]. The transition matrix indicates the probability of the pavement being in one state and the probability of transition from the current state to another state of deterioration.

Figure 6 shows an example of a Markovian representation of deterioration of pavements. The circles represent the pavement states, and the arrows indicate the possible pavement deterioration states in each year. The transition probability associated with each arrow indicates the probability of deterioration between two states [49]. As can be observed, pavements can deteriorate from *state 1* at $t = 0$ to *states 1, 2, and 3* at $t = 1$.

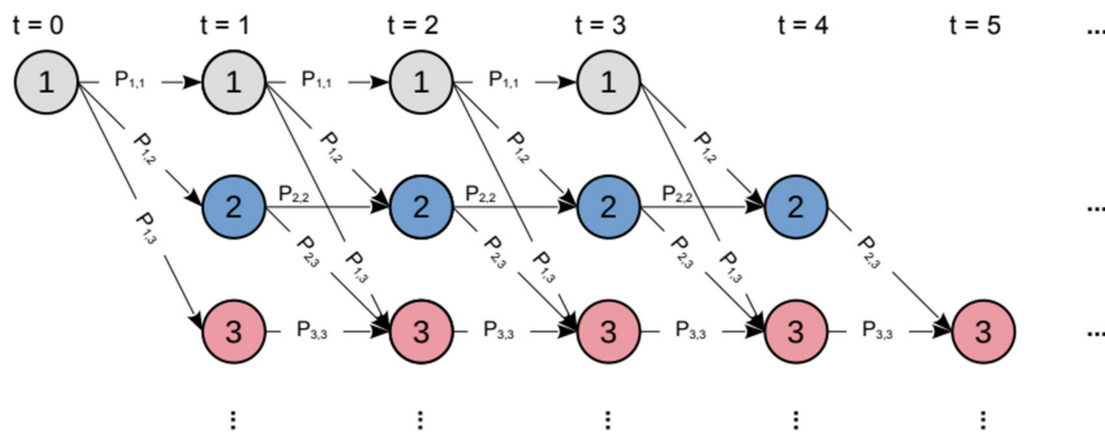


Figure 6. Markovian representation of deterioration (adapted from [49]).

Briefly, the Markov diagram is no more than an enumeration of all possible pavement deterioration sequences with associated transition probabilities. These probabilities can also be interpreted as the expected proportions of pavements in each state.

$P_{i,j}$ represents the transition probability from *state i* to *state j* when no M&R action is applied to pavements. In this case, the pavement changes to a worse state, and to move to a better state, it is necessary to apply maintenance actions. As the Markov is time independent, the transition probability from *state i* to *state j* in year *t* is equal to the transition probability from *state i* to *state j* in year *t + 1* [49].

The first probabilistic performance model was presented by [50], whereas the first modern network-level PMS was developed for the Arizona Department of Transportation [51].

Wang et al. [52] presented a methodology for calculating the transition probabilities using the pavement miles that transit from one state to another.

Li et al. [45] discussed the development of a non-homogeneous Markov probabilistic program for modeling pavement deterioration. A non-homogeneous probability model is defined in terms of states, stages, and a sequence of transition matrices.

Ferreira et al. [53] presented a segment-linked optimization model (see Equation (11)) to be used within PMSs. It allows M&R actions to be defined for specific segments of a road network, overcoming one of the principal drawbacks of the widely used Arizona PMS (the absence of an explicit spatial dimension).

$$\sum_{k=1}^K x_{s,j,k,t} = \sum_{k=1}^K \sum_{i=1}^I x_{s,j,k,t-1} P_{i,j,k} \quad (11)$$

where:

S is the number of road segments;

$x_{s,j,k,t}$ is the proportion of pavement of segment *s* in state *j* at the beginning of period *t* to which action *k* is applied; and

$P_{i,j,k}$ is the transition probability from state *i* to state *j* when action *k* is applied to the pavement.

Mishalani and Madanat [54] developed a probabilistic-based model to estimate the transition probabilities based on the time spent (duration) in a given state.

Other researchers have developed methods that minimize the sum of residuals (errors), defined as the difference between the observed distress ratings and their corresponding predicted values obtained from the Markov model [44,55,56].

Madanat et al. [57] revealed that one of the most common methods for estimating transition probabilities is the expected-value method. In this method, the data is first divided into similar behavior groups with similar attributes. A linear regression model is

fitted for each group, with the condition rating as the dependent variable and age t as the independent variable. A transition matrix is then estimated for each group by minimizing the distance between the expected value of the condition rating obtained from the linear regression model and the theoretically expected value derived from the Markov chain structure.

Yang et al. [34] used a dynamic or recurrent Markov chain for modeling pavement crack deterioration and a logistic model to calculate the transition probability matrix.

Pulugurta et al. [58] developed a first-order homogenous Markov model to forecast pavement distresses and PCR using the Ohio Department of Transportation (ODOT) database. Each distress was divided into different states based on their severity and extent.

Abaza and Murad [48] developed a stochastic approach to estimate the required design thickness for flexible pavement using typical design factors and new additional stochastic-based factors. The long-term performance of pavement has been traditionally defined using a pavement performance curve. The discrete-time Markov model typically applies the transition probabilities (transition matrix) and the initial state probabilities to predict the future pavement distress ratings over an analysis period. The predicted pavement distress ratings are used to construct the corresponding performance curve. The transition probabilities along the matrix main diagonal $P_{i,i}$ represent the probability that pavements presently in *state* i will remain in the same condition state after the elapse of one transition.

$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & 0 & 0 & \cdots & 0 \\ 0 & P_{2,2} & P_{2,3} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & P_{m-1,m-1} & P_{m-1,m} \\ 0 & 0 & 0 & 0 & 0 & P_{m,m} \end{pmatrix}$$

where: $P_{i,i} + P_{i,i+1} = 1.0$ and $P_{m,m} = 1.0$.

The transition probabilities $P_{i,i+1}$ represent pavement deterioration rates from a present state i to a worse state $i + 1$ after one transition. All matrix entries below the main diagonal represent pavement improvement rates, which are assigned zero values in the absence of M&R works. The main objective in defining the transition matrix is to predict the future pavement conditions of new pavements.

Several researchers have extensively used the discrete-time Markov model to predict pavement performance [43–47]. Predicted pavement performance curves were used to yield optimum pavement and estimate overlay design thickness [47,59,60].

8.3.2. The Nonhomogeneous Markov Process

The nonhomogeneous Markov process assumes transition probabilities are time-dependent and are more consistent with reality since traffic (volume, growth rate, truck percentage) and environmental conditions (temperature, precipitation) vary throughout the period of analysis [45]. Consequently, the pavement deterioration process is defined by a sequence of transition probability matrices (TPM) represented by Equation (12).

$$\sum_{k=1}^K w_{j,k,t} = \sum_{i=1}^I \sum_{k=1}^K w_{i,k,t-1} P_{i,j,k} \quad j = 1, \dots, J; \quad t = 2, \dots, T \quad (12)$$

where: $P_{i,j,k,t-1}$ is the transition probability from state i to state j , during year $t - 1$, when action k is applied to pavements.

The transition probability from state i to state j in year t is different than the transition probability from state i to state j in year $t + 1$.

The application of the nonhomogeneous Markov process to network level pavement management requires the computation of transition probability matrices for each year of the period of analysis, which implies a significant increase in the problem size.

Li et al. [45] calculated TPM for a pavement section located in Canada, where each element of the TPM was determined using the Monte Carlo simulation technique.

In the research by [45,52], it was considered that the pavement performance degradation could be modeled using the nonhomogeneous (i.e., non-stationary) discrete Markov chain, i.e., a Markov process with discrete natural parameters and discrete state space.

Hong and Wang [61] developed a probabilistic approach for predicting pavement performance based on a nonhomogeneous continuous Markov chain.

8.3.3. The Semi-Markov Process

The semi-Markov process is a variant of the homogeneous Markov process for which time is not fixed. A pavement that is in a given state deteriorates to another in a period of time that is variable and follows a probabilistic distribution. Pavement deterioration can be illustrated conceptually as a function of time or traffic. The semi-Markov process is motivated by the desire to exploit a vital timing element of this discrete approximation of pavement deterioration.

Figure 7 shows a semi-Markov representation of deterioration with fixed holding time. The number of years associated with the arrows indicates the amount of time the pavements remain in each state.

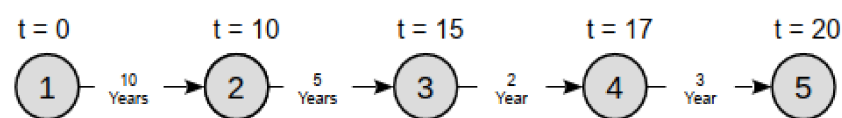


Figure 7. Semi-Markov representation of deterioration with fixed holding time (adapted from [49]).

The semi-Markov process is based on the idea of assigning a holding time to each pavement state. Therefore, to find out what state pavements will be in at time t , all that is required is to move t time periods to the right [49]. To consider several pavements, rather than assigning a fixed holding time to each arrow, a probability distribution over holding time is assigned. Figure 8 shows the semi-Markov representation of maintenance and rehabilitation.

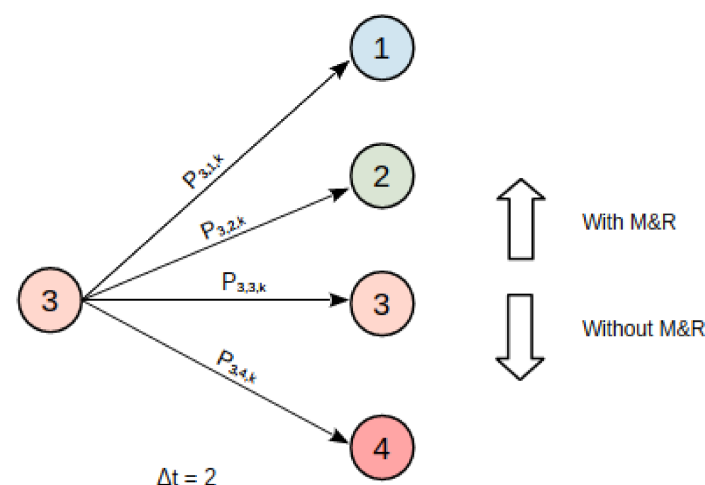


Figure 8. Semi-Markov representation of maintenance and rehabilitation.

In the semi-Markov process, M&R actions can be represented by multiple arrows that point to other states, just as in the Markov process. At the same time, a probability distribution over holding time is assigned to each arrow. $P_{i,j,k}$ is the probability that a pavement that has just entered state i will transition to state j when action k is applied. $H_{i,j,k,t}$ is the probability distribution over the time the pavement will remain in state i before it makes the transition to state j when action k is applied. The holding time distribution

can be interpreted as a histogram over the time it takes to deteriorate from one state to another [49].

Within the semi-Markov process, the number of time periods is reduced, but other data are necessary, namely, the holding time distributions for each state. The application of the semi-Markov process in pavement management at the network level is difficult due to the existence of a transition interval matrix between states, which is not equal for all the pavements in the network.

Moghaddass et al. [62] investigated how interval-censored inspection data can be used for parameter estimation and reliability analysis of a multi-state device. A general stochastic process called a non-homogeneous continuous-time semi-Markov process (NHCTSM) was considered for the degradation process, which has the flexibility to cover many of the previously studied degradation models used in the literature.

9. Conclusions, Discussion, and Guidelines to Support the Development of PPPMs

The main goal of this article was to review the most common techniques and provide guidelines to support the development of pavement performance prediction models.

It is important to know that some essential aspects need to be considered when developing PPPMs, as illustrated in Figure 9.

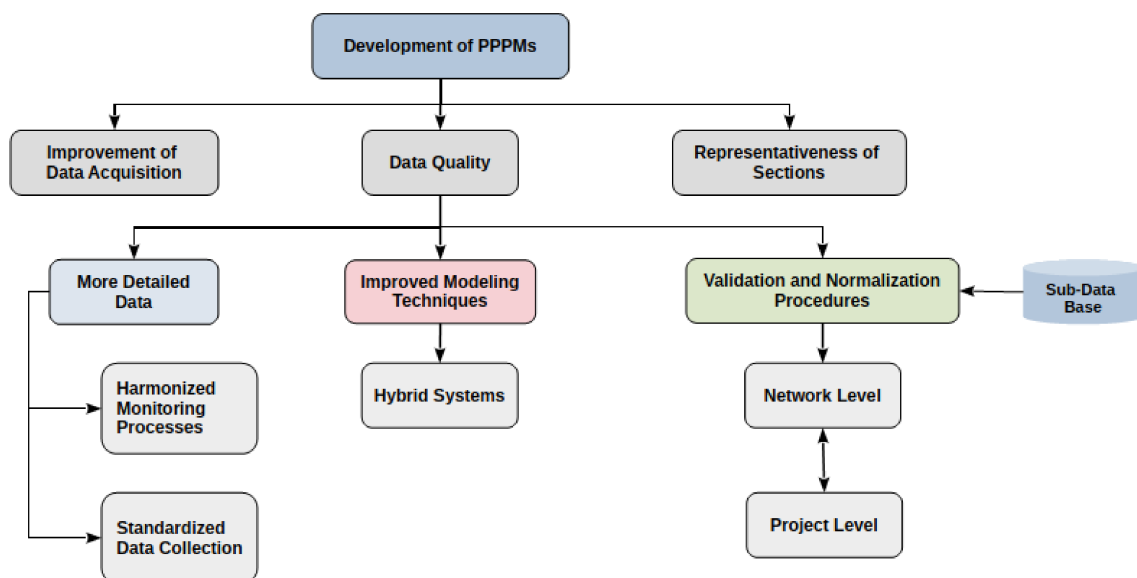


Figure 9. Important aspects in the development of PPPMs.

First, the sample of sections used in the development of the models must represent the type of pavement, have a wide range of ages, and be relevant to the network in question.

Secondly, the quality of the input data is crucial for the final fitting of the models, and, consequently, the expected result will be more accurate and adjusted to reality. Therefore, it is essential to improve data acquisition through standardized data collection methods and harmonized monitoring processes.

Another important aspect is enhancing the connection between the models developed and the rest of the network. This connection can be made with a subset of the original database. For this subset of the database, non-destructive and destructive tests, such as the falling weight deflectometer test, can be prepared to validate the information about the pavement structure (structural number, the thickness of the layers, and CBR data). Consequently, the introduction of structural parameters into the analysis will be possible.

Finally, it is crucial to use improved modeling techniques (ML algorithms), and when new data is available, the models should be updated. Machine learning modeling

techniques are essential in the presence of large amounts of data, which represents one of the challenges faced by road agencies.

In summary, machine learning is a set of programming techniques that aim to find patterns in data to perform future predictions. Statistics are used in machine learning to build mathematical models because the core task is making inferences from a sample. Its parameters define a model, and “learning” is the process of optimizing the model’s parameters using the training data. Then, the model is tested with a new test data set to validate the model’s prediction capabilities. The final model may be predictive to make predictions about future events, descriptive to gain knowledge from data, or both. Different considerations need to be taken into account depending on the type of model under development. For prediction models, the best model is the one that provides the lowest misclassification rate for both training and testing data sets.

Choosing the right machine learning algorithm is overwhelming since each takes a different approach to learning, and there are plenty of algorithms that can be selected. Therefore, finding the right algorithm can be considered a trial-and-error process. However, having a clear vision about the size and type of data to be worked with, the insights to be extracted from the data, and how they will be used will help narrow down the machine learning algorithms list.

In terms of evaluating the models, it is crucial to ensure that the data used to develop the models are not the same as the ones used in the evaluation. Several sampling methods allow data to be divided and help to avoid overfitting:

- Hold-out test set—divides data into a training set and a testing set;
- Hold-out sampling—divides data into a training set, a validation set, and a test set;
- *k*-Fold cross validation—data are divided into *k* equal-size folds. The first fold is used as a test set, and the remaining *k* – 1 folds as training sets. The process is repeated for all *k* folds;
- Leave-one-out cross validation—*k*-fold cross-validation in which the number of folds is the same as the number of training instances;
- Bootstrapping—preferred over cross-validation for small data sets;
- Out-of-time sampling—a hold-out sampling that is targeted rather than random.

Typical performance measures to assess the quality of the final model are the misclassification rate (Equation (13)), the confusion matrix, and the classification accuracy.

$$\text{Misclassification Rate} = \frac{\text{N}^\circ \text{ of Incorrect Predictions of the Model}}{\text{N}^\circ \text{ of Total Predictions of the Model}} \quad (13)$$

The confusion matrix calculates the frequencies of each possible outcome of the model predictions, so for binary problems, the confusion matrix has four possible outcomes:

- True positive (TP);
- True negative (TN);
- False positive (FP);
- False negative (FN).

The first two correspond to correct predictions made by the model and the last two to incorrect ones.

The classification accuracy is the inverse of the misclassification rate and is defined in Equation (14).

$$\text{Classification Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (14)$$

Nonetheless, a trade-off between specific characteristics of the algorithms must be considered. Those characteristics are shown in Figure 10 and specified briefly in Table 6.

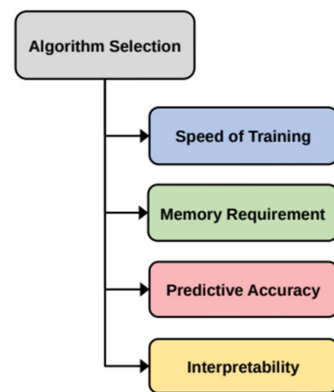


Figure 10. Algorithm selection criteria.

Table 6. Trade-off characteristics of some machine learning algorithms.

| Type of Algorithm | Prediction Speed | Training Speed | Memory Usage | Required Tuning | General Assessment |
|---------------------------------------|------------------|----------------|-----------------|-----------------|--|
| Logistic Regression + Linear SVM | Fast | Fast | Small | Minimal | Suitable for small problems with linear decision boundaries |
| Decision Trees | Fast | Fast | Small | Some | Good generalist but prone to overfitting |
| Nonlinear SVM and Logistic Regression | Slow | Slow | Medium | Some | Suitable for many binary problems and handles high-dimensional data well |
| Nearest Neighbor | Moderate | Minimal | Medium | Minimal | Lower accuracy, but easy to use and interpret |
| Naïve Bayes | Fast | Fast | Medium | Some | Widely used for text, including spam filtering |
| Ensembles | Moderate | Slow | Varies | Some | High accuracy and good performance for small to medium-sized data sets |
| Neural Network | Moderate | Slow | Medium to Large | Lots | Popular for classification, compression, recognition, and forecasting |

It is important to remember that almost every approach can work for both continuous and categorical descriptive and target features. However, specific techniques are a more natural fit for some data than others (see Figure 11). The first thing to consider about data is whether the target feature is continuous or categorical.

In many cases, data sets will contain both categorical and continuous descriptive features. The most naturally suited learning approaches in these scenarios are probably those that are best suited to the majority feature type.

The last issue to consider concerning data when selecting machine learning approaches is the curse of dimensionality. If there are a large number of descriptive features, then we will need a large training data set. Feature selection is an essential process in any machine learning project and should generally be applied no matter what type of models are being developed.

To conclude, this article addressed one of the main concerns in pavement management, which is predicting the future condition of the road network. The next goal is being able to select and apply the M&R operations within the agency's available budget and linking the project-level and network-level decisions.

In future research, this topic related to decision-making for optimizing the M&R selection will be covered using reinforcement learning models.

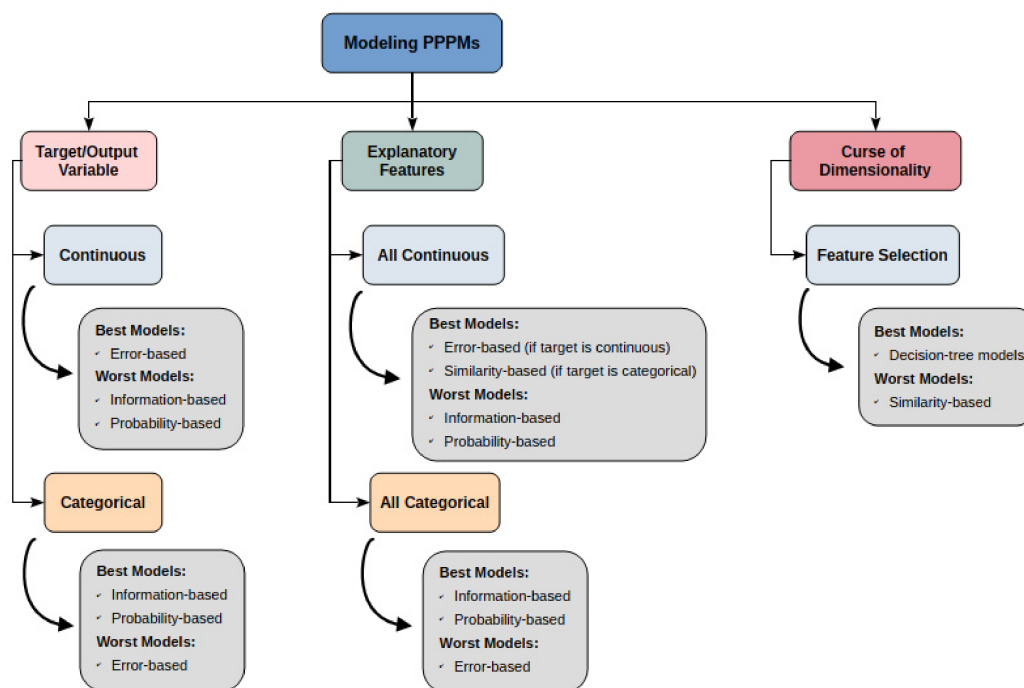


Figure 11. PPPM considerations for model selection according to data.

Author Contributions: Conceptualization, R.J.-S., A.F., and G.F.; methodology, R.J.-S., A.F., and G.F.; investigation, R.J.-S.; writing—original draft preparation, R.J.-S.; writing—review and editing, R.J.-S., A.F., and G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author Rita Justo-Silva is grateful to the Portuguese Foundation for Science and Technology for her MIT-Portugal grant (PD/BD/113721/2015).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|-------|--|
| PPPMs | Pavement performance prediction models |
| PMSs | Pavement management systems |
| ML | Machine learning |
| SL | Supervised learning |
| UL | Unsupervised learning |
| RL | Reinforcement learning |

References

1. Branco, F.; Pereira, P.; Picado-Santos, L. *Pavimentos Rodoviários*; Edições Almedina: Coimbra, Portugal, 2016.
2. Yang, J.; Lu, J.J.; Gunaratne, M. *Application of Neural Network Models for Forecasting of Pavement Crack Index and Pavement Condition Rating*; Technical Report; University of South Florida: Tampa, FL, USA, 2003.
3. Odoki, J.B.; Kerali, H.R. *HDM-4 Volume 4: Analytical Framework and Model Descriptions*; World Road Association (PIARC): Paris, France, 2013.
4. OCDE. *Essai OCDE en Vraie Grandeur des Superstructures Routières*. In *Recherche en Matière de Routes et de Transports Routiers*; Organisation de Coopération et de Développement Économiques (OCDE): Paris, France, 1991.
5. COST324. *Long Term Performance of Road Pavements*; Final Report of the Action; European Commission: Paris, France, 1997.

6. *Pavement Deterioration Models: Deliverable D4-RO 96-SC.404*; Technical Report; European Commission: Paris, France, 1998.
7. MathWorks. Introducing Machine Learning. Available online: <https://www.mathworks.com/content/dam/mathworks/ebook/gated/machineLearning-ebook.pdf> (accessed on 18 July 2019).
8. Kelleher, J.D.; Mac Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*; MIT Press: Cambridge, MA, USA, 2015.
9. AASHTO. *Pavement Management Guide*, 2nd ed.; Technical Report; American Association of State Highway and Transportation Officials: Washington, DC, USA, 2012.
10. Sadek, A.W.; Freeman, T.E.; Demetsky, M.J. Deterioration prediction modeling of Virginia's interstate highway system. *Transp. Res. Rec.* **1996**, *1524*, 118–129. [[CrossRef](#)]
11. Haider, S.W.; Chatti, K. Effect of design and site factors on fatigue cracking of new flexible pavements in the LTPP SPS-1 experiment. *Int. J. Pavement Eng.* **2009**, *10*, 133–147. [[CrossRef](#)]
12. Kaur, D.; Pulugurta, H. Comparative analysis of fuzzy decision tree and logistic regression methods for pavement treatment prediction. *WSEAS Trans. Inf. Sci. Appl.* **2008**, *5*, 979–990.
13. Henning, T.F.P.; Costello, S.B.; Watson, T.G. *A Review of the HDM/dTIMS Pavement Models Based on Calibration Site Data*; Land Transport New Zealand: Auckland, New Zealand, 2006.
14. Li, Z. A Probabilistic and Adaptive Approach to Modeling Performance of Pavement Infrastructure. Ph.D. Thesis, University of Texas, Austin, TX, USA, 2005.
15. Ben-Akiva, M.; Humplick, F.; Madanat, S.; Ramaswamy, R. Infrastructure management under uncertainty: Latent performance approach. *J. Transp. Eng.* **1993**, *119*, 43–58. [[CrossRef](#)]
16. Ben-Akiva, M.; Gopinath, D. Modeling infrastructure performance and user costs. *J. Infrastruct. Syst.* **1995**, *1*, 33–43. [[CrossRef](#)]
17. Chu, C.Y.; Durango-Cohen, P.L. Estimation of infrastructure performance models using state-space specifications of time series models. *Transp. Res. Part C Emerg. Technol.* **2007**, *15*, 17–32. [[CrossRef](#)]
18. Durbin, J.; Koopman, S.J. *Time Series Analysis by State Space Methods*; Oxford University Press: Oxford, UK, 2012.
19. Durango-Cohen, P.L. A time series analysis framework for transportation infrastructure management. *Transp. Res. Part B Methodol.* **2007**, *41*, 493–505. [[CrossRef](#)]
20. Lorino, T.; Lepert, P.; Marion, J.M.; Khraibani, H. Modeling the road degradation process: Nonlinear mixed effects models for correlation and heteroscedasticity of pavement longitudinal data. *Procedia Soc. Behav. Sci.* **2012**, *48*, 21–29. [[CrossRef](#)]
21. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974. [[CrossRef](#)]
22. Ware, J.H. Linear models for analysis of longitudinal studies. *Am. Stat.* **1985**, *39*, 95–101.
23. Diggle, P.J.; Liang, K.L.; Zeger, S.L. *Analysis of Longitudinal Data*; Oxford University Press: New York, NY, USA, 1994.
24. Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurement Data*; Chapman and Hall: London, UK, 1995.
25. Vonesh, E.; Chinchilli, V.M. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*; CRC Press: Boca Raton, FL, USA, 1996.
26. Ricardo Archilla, A.; Madanat, S. Statistical model of pavement rutting in asphalt concrete mixes. *Transp. Res. Rec.* **2001**, *1764*, 70–77. [[CrossRef](#)]
27. Karballaezadeh, N.; Mohammadzadeh, S.D.; Shamsirband, S.; Hajikhodaverdikhan, P.; Mosavi, A.; Chau, K.W. Prediction of remaining service life of pavement using an optimized support vector machine (case study of Semnan–Firuzkuh road). *Eng. Appl. Comput. Fluid Mech.* **2019**, *13*, 188–198. [[CrossRef](#)]
28. Ziari, H.; Maghrebi, M.; Ayoubinejad, J.; Waller, S.T. Prediction of pavement performance: Application of support vector regression with different kernels. *Transp. Res. Rec.* **2016**, *2589*, 135–145. [[CrossRef](#)]
29. Wasserman, P.D. *Neural Computing*; Van Nostrand Reinhold: New York, NY, USA, 1989; pp. 44–54.
30. Huang, Y.; Moore, R. Roughness level probability prediction using artificial neural networks. *Transp. Res. Rec.* **1997**, *1592*, 89–97. [[CrossRef](#)]
31. Flintsch, G.W.; Zaniewski, J.P. Expert project recommendation procedure for Arizona Department of Transportation's pavement management system. *Transp. Res. Rec.* **1997**, *1592*, 26–34. [[CrossRef](#)]
32. Bosurgi, G.; Trifirò, F.; Xilbilia, M.G. Artificial neural network for predicting road pavement conditions. In Proceedings of the 4th International SIV Congress, Palermo, Italy, 12–14 September 2007.
33. Prozzi, J.A.; Madanat, S.M. Incremental nonlinear model for predicting pavement serviceability. *J. Infrastruct. Syst.* **2003**, *129*, 635–641. [[CrossRef](#)]
34. Yang, J.; Gunaratne, M.; Lu, J.J.; Dietrich, B. Use of recurrent Markov chains for modeling the crack performance of flexible pavements. *J. Transp. Eng.* **2005**, *131*, 861–872. [[CrossRef](#)]
35. Kononenko, I.; Kukar, M. *Machine Learning and Data Mining*; Horwood Publishing: Cambridge, UK, 2007.
36. Smith, W.; Finn, F.; Kulkarni, R.; Saraf, C.; Nair, K. NCHRP Report 213: Bayesian methodology for verifying recommendations to minimize asphalt pavement distress. In *Transportation Research Board, National Research Council*; Transportation Research Board: Washington, DC, USA, 1979.
37. Haper, W.V.; Majidzadeh, K. Utilization of expert opinion in the two pavement management systems. In Proceedings of the 70th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 15 January 1991.
38. Hajek, J.J.; Bradbury, A. Pavement performance modeling using Canadian strategic highway research program bayesian statistical methodology. *Transp. Res. Rec.* **1996**, *1524*, 160–170. [[CrossRef](#)]

39. Pereira, P.A.; Barbosa, N. *Sistema de Gestão da Conservação—Manual de Utilização do PRISM*; Universidade do Minho: Braga, Portugal, 1998.
40. Hong, F.; Prozzi, J.A. Updating pavement deterioration models using the Bayesian principles and simulation techniques. In Proceedings of the 1st Annual Inter-University Symposium on Infrastructure Management, Waterloo, ON, Canada, 6 August 2005.
41. Jiménez, L.A.; Mrawira, D. Bayesian regression in pavement deterioration modeling: Revisiting the AASHTO road test rut depth model. *Infrastruct. Vial* **2012**, *14*, 28–35.
42. Litao, L. A Methodology for Developing Performance-related Specifications for Pavement Preservation Treatments. Ph.D. Thesis, Texas A&M Transportation Institute, Bryan, TX, USA, 2013.
43. Way, G.; Eisenberg, J.; Kulkarni, R. Arizona Pavement Management System: Phase 2, Verification of performance prediction models and development of data base. *Transp. Res. Rec.* **1982**, *846*, 49–55.
44. Butt, A.; Shahin, M.; Feighan, K.; Carpenter, S. Pavement performance prediction model using the Markov process. *Transp. Res. Rec.* **1987**, *1123*, 12–19.
45. Li, N.; Xie, W.C.; Haas, R. Reliability-based processing of Markov chains for modeling pavement network deterioration. *Transp. Res. Rec.* **1996**, *1524*, 203–213. [[CrossRef](#)]
46. Abaza, K.; Murad, M. Dynamic probabilistic approach for long-term pavement restoration program with added user cost. *Transp. Res. Rec.* **2007**, *1990*, 48–56. [[CrossRef](#)]
47. Abaza, K.; Murad, M. Predicting flexible pavement remaining strength and overlay design thickness with stochastic modeling. *Transp. Res. Rec. J. Transp. Res. Board* **2009**. [[CrossRef](#)]
48. Abaza, K.; Murad, M. Stochastic approach for design of flexible pavement. *Road Mater. Pavement Des.* **2011**, *12*, 663–685.
49. Ferreira, A.; Picado-Santos, L.; Antunes, A. Pavement performance modelling: State of the art. In Proceedings of the Seventh International Conference on Civil and Structural Engineering Computing, Egmond aan Zee, The Netherlands, 13–15 September 1999; pp. 157–264.
50. Karan, M.A.; Haas, R.C.G. Determining investment priorities for urban pavement improvements. *Proc. Assoc. Asph. Paving Technol.* **1976**, *45*, 254–282.
51. Golabi, K.; Kulkarni, R.B.; Way, G.B. A Statewide pavement management system. *Interfaces* **1982**, *12*, 5–21. [[CrossRef](#)]
52. Wang, K.C.P.; Zaniewski, J.; Way, G. Probabilistic behavior of pavements. *J. Transp. Eng.* **1994**, *120*, 358–375. [[CrossRef](#)]
53. Ferreira, A.; Picado-Santos, L.; Antunes, A. A segment-linked optimization model for deterministic pavement management systems. *Int. J. Pavement Eng.* **2002**, *3*, 95–102. [[CrossRef](#)]
54. Mishalani, R.; Madanat, S. Computation of infrastructure transition probabilities using stochastic models. *J. Infrastruct. Syst.* **2002**, *8*, 139–148. [[CrossRef](#)]
55. Shahin, M.Y. *Pavement Management for Airports, Roads, and Parking Lots*; Springer: New York, NY, USA, 2005; Volume 501.
56. Ortiz-Garcia, J.; Costello, S.; Snaith, M. Derivation of transition probability matrices for pavement deterioration. *J. Transp. Eng.* **2006**, *132*, 141–161. [[CrossRef](#)]
57. Madanat, S.; Mishalani, R.; Ibrahim, W.H.W. Estimation of infrastructure transition probabilities from condition rating data. *J. Infrastruct. Syst.* **1995**, *1*, 120–125. [[CrossRef](#)]
58. Pulugurta, H.; Shao, Q.; Chou, Y.J. Pavement condition prediction using Markov process. *J. Stat. Manag. Syst.* **2009**, *12*, 853–871. [[CrossRef](#)]
59. Abaza, K.; Abu-Eisheh, S. An optimum design approach for flexible pavement. *Int. J. Pavement Eng.* **2003**, *4*, 1–11. [[CrossRef](#)]
60. Abaza, K. Performance-based models for flexible pavement structural overlay design. *J. Transp. Eng.* **2005**, *131*, 149–159. [[CrossRef](#)]
61. Hong, H.P.; Wang, S.S. Stochastic modeling of pavement performance. *Int. J. Pavement Eng.* **2003**, *4*, 235–243. [[CrossRef](#)]
62. Moghaddass, R.; Zuo, M.J.; Liu, Y.; Huang, H.Z. Predictive analytics using a nonhomogeneous semi-Markov model and inspection data. *IIE Trans.* **2015**, *47*, 505–520. [[CrossRef](#)]