UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITÀ DI PISA

Technische Universität Braunschweig

International Doctorate in Civil and Environmental Engineering

Infrastructures Resilience: Geostatistical Analyses and Machine Learning Techniques for Optimizing Infrastructural Maintenance and Safety Level

PhD Student: Nicholas Fiorentini

Info	
Home Institution	Department of Civil and Industrial Engineering, Engineering School of the University of Pisa, Largo Lucio Lazzarino 1, 56126, Pisa, Italy
Joint Supervision	Institute of Geodesy and Photogrammetry, Technische Universität Braunschweig, Bienroder Weg 81, 38106, Braunschweig, Germany
Italian Tutor/Co Tutor	Professor Massimo Losa
Foreign Tutor/Co Tutor	Professor Markus Gerke
Email	nicholas.fiorentini@phd.unipi.it; nicholas.fiorentini@unifi.it;

Abstract

This Ph.D. thesis proposes a compendium of researches in the field of road pavement and road safety management. The primary objective is to provide the road authority with some operational tools capable of providing support in decision-making processes concerning several activities, such as infrastructure monitoring, infrastructure management, and infrastructure planning.

These tools steam from advanced statistics and data science algorithms, which allow extrapolating information and meaningful patterns through observation and inference from large amounts of data.

First part
Road Pavement Management: Calibration of Machine Learning Algorithms for surface motion prediction beneath road pavement structure

Technologically advanced strategies in infrastructural maintenance are increasingly required in countries such as Italy, where recovery and rehabilitation interventions are preferred to new works. For this purpose, Interferometric Synthetic Aperture Radar (InSAR) techniques have been employed in recent years, achieving reliable outcomes in the identification of infrastructural instabilities. Nevertheless, using the InSAR survey exclusively, it is not feasible to recognize the reasons for such vulnerabilities, and further in-depth investigations are essential. The primary purpose of this first step of the Ph.D. is to introduce a methodology for predicting and mapping surface motion beneath road pavement structures

caused by environmental factors (such as subsidence). Persistent Scatterer InSAR (PS-InSAR) measurements, geospatial analyses, and Machine Learning Algorithms (MLAs) are employed for achieving the purpose. Two single learners, i.e., Regression Tree (RT) and Support Vector Machine (SVM), and two ensemble learners, i.e., Boosted Regression Trees (BRT) and Random Forest (RF) are employed for estimating the surface motion ratio in terms of mm/year over the Province of Pistoia (Tuscany Region, central Italy, 964 km^2), in which strong subsidence phenomena have occurred. The interferometric process of 210 Sentinel-1 images from 2014 to 2019 allows exploiting the average displacements of 52,257 Persistent Scatterers as output targets to predict. A set of 29 environmental-related factors are preprocessed by GIS platforms and employed as input features. Once the dataset has been prepared, three wrapper feature selection approaches (backward, forward, and bi-directional) are used for recognizing the set of most relevant features to be used in the modeling. A random splitting of the dataset in 70% and 30% is implemented to identify the training and test set. Through a Bayesian Optimization Algorithm (BOA) and a 10-Fold Cross-Validation (CV), the algorithms are trained and validated. Therefore, the Predictive Performance of MLAs is evaluated and compared by plotting the Taylor Diagram. Outcomes show that SVM and BRT are the most suitable algorithms; in the test phase, BRT has the highest Correlation Coefficient (0.96) and the lowest Root Mean Square Error (0.44 mm/year), while the SVM has the lowest difference between the standard deviation of its predictions (2.05 mm/year) and that of the reference samples (2.09 mm/year). Finally, algorithms are used for mapping surface motion over the study area. This first section concludes with the discussion of three case studies on critical stretches of two-lane rural roads for evaluating the reliability of the procedure. Road authorities could consider the proposed tools for their monitoring, management, and planning activities.

Second Part
Road Pavement Management: In-Situ Validation of the MLAs

This second part of the Ph.D. argues an expeditious but reliable methodology for identifying critical road sites that need attention by the road authorities. Starting from the outcomes of the first part, the aim is to integrate the information provided by MLAs with profilometric surface roughness measurements and relative calculation of the International Roughness Index (IRI). The integration phase consists of the following steps: (1) carrying out profilometric measurements, (2) computing the IRI, (3) making prediction by MLAs and resampling on sections of 100-meters in length, and (4) comparing outcomes. Where a dependence between the two measurements exists, MLAs could replace the IRI surveys, with considerable savings in time and costs for the road authority. Contrarily, in the event that the dependence is not detected, it is possible to conclude that both measurements are essential for the correct management of road pavements. Outcomes show that an in-situ measurement cannot be completely replaced by a model prediction, but the latter is effective in the most serious situations. Therefore, once the calibrated MLAs have been exploited, Road authorities will be able to conduct targeted inspections on a limited set of road sites, significantly saving time and money.

Third Part
Road Safety Management: Development of Crash Prediction Models

Crash Prediction Models (CPMs) can be reliable tools able to identify road safety issues and prevent severe accidents even in large and complex road network. In this third step of the Ph.D., the goal is to develop several Negative Binomial-based CPMs in the context of Italian two-lane roads. Five types of CPM have been developed for predicting the number of crashes per year (Fatal and Injuries crashes): (1) Base model (B-CPM), that is a CPM with intercept and traffic flow effects only, (2) Multivariate model (M-CPM), that is a CPM with main effects of intercept, traffic flow, driveway density, intersection density, slope, horizontal and vertical alignment, width of the lanes, and road context (urban, suburban, rural), (3) Locally Optimized Multivariate model (LOM-CPM), that is a specific M-CPM for the analyzed network, in which a specific factor to account for potential intrinsic effects of each different road has been added, and (4-5) Main and Interaction Effect (MIE) models, both for M-CPM (MIEM-CPM) and LOM-CPM (MIELOM-CPM), that are CPMs in which models can consider the effect of interactions between features. The analyzed network is managed by the Tuscany Region Road Administration (TRRA), Central Italy, and extends for about 1,000 km. It has been investigated an amount of 1818 road stretches of 500 meters in length in which 5802 Fatal and Injuries crashes occurred from year 2008 to 2016. The Goodness-of-Fit (GoF) assessment of the CPMs has been carried by the Log-Likelihood (LL), Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Correlation Coefficient (R), Determination Coefficient (R^2), Scatterplot, and Cumulative Residual (CURE) plot. The GoF analysis demonstrates the reliability of the CPMs' predictions. B-CPM and M-CPM can be used for expeditious safety analyses and for predicting crash frequency in other networks that share the application field of the developed CPMs. The LOM-CPM appears more suitable than the simpler M-CPM to be used by TRRA for safety analyses. Finally, the MIELOM-CPM and MIEM-CPM outperform the LOM-CPM and M-CPM, respectively, demonstrating that the contribution of interaction effects deserves attention and should be accounted in road safety analyses. Nonetheless, they are also the more complex and less interpretable CPMs.

Fourth Part
Road Safety Management: Developing of a Road Screening Procedure

Screening procedures in road blackspot detection are essential tools for road authorities for quickly gathering insights on the safety level of each road site they manage. The fourth phase of the Ph.D. suggests a road blackspot screening procedure for two-lane rural roads, relying on five different MLAs and real long-term traffic data. The network analyzed is the one managed by the TRRA, mainly composed of two-lane rural roads. An amount of 995 road sites, where at least one accident occurred in 2012–2016, have been labeled as "Accident Case". Accordingly, an equal number of sites where no accident occurred in the same period, have been randomly selected and labeled as "Non-Accident Case". Five different MLAs, namely Logistic Regression, Classification and Regression Tree, Random Forest, K-Nearest Neighbor, and Naïve Bayes, have been trained and validated. The output response of the MLAs, i.e., crash occurrence susceptibility, is a binary categorical variable. Therefore, such algorithms aim to classify a road site as likely safe ("Accident Case") or potentially susceptible to an accident occurrence ("Non-Accident Case") over five years. Finally, algorithms have been compared by a set of performance metrics, including precision, recall, F1-score, overall accuracy, confusion matrix, and the Area Under the Receiver Operating Characteristic. Outcomes show that the Random Forest outperforms the other MLAs with an overall accuracy of 73.53%. Furthermore, all the MLAs do not show overfitting issues. Road authorities could consider MLAs to draw up a priority list of on-site inspections and maintenance interventions.

Fifth Phase
Road Safety Management: Calibrating MLAs for Crash Severity Prediction

Crash severity is undoubtedly a fundamental aspect of a crash event. Although MLAs for predicting crash severity have recently gained interest by the academic community, there is a significant trend towards neglecting the fact that crash datasets are acutely imbalanced. Overlooking this fact generally leads to weak classifiers for predicting the minority class (crashes with higher severity). In this fifth step of the Ph.D., in order to handle imbalanced accident datasets and provide a better prediction for the minority class, the random undersampling the majority class (RUMC) technique is used. By employing an imbalanced and a RUMC-based balanced training set, the calibration, validation, and evaluation of four different crash severity predictive models, including random tree, k-nearest neighbor, logistic regression, and random forest, have been proposed. Accuracy, true positive rate (recall), false positive rate, true negative rate, precision, F1-score, and the confusion matrix have been calculated to assess the performance. Outcomes show that RUMC-based models provide an enhancement in the reliability of the classifiers for detecting fatal crashes and those causing injury. Indeed, in imbalanced models, the true positive rate for predicting fatal crashes and those causing injury spans from 0% (logistic regression) to 18.3% (k-nearest neighbor), while for the RUMC-based models, it spans from 52.5% (RUMC-based logistic regression) to 57.2% (RUMC-based k-nearest neighbor). Organizations and decision-makers could make use of RUMC and MLAs in predicting the severity of a crash occurrence, managing the present, and planning the future of their works.