## DOTTORATO DI RICERCA IN INFRASTRUTTURE E TRASPORTI

## SCHEDA PER L'AMMISSIONE AL III ANNO DI CORSO

|  |  |  |  |
|---|---|---|---|
| **Dottorando** | Ken Koshy Varghese | **Ciclo** | XXXVII |
| **Curriculum** | Transport and Territory Planning | **Tutore** | Prof. Guido Gentile |

**Argomento della ricerca** Advanced Machine Learning Approaches for Time Series Analysis in Transportation: Conformal Prediction, Uncertainty Quantification, and Interpretable Frameworks.

### SEZIONE A

**Ricerca di Dottorato**

**(massimo 5 pagine)**

**1 – Aggiornamento del programma logico e cronologico delle attività** *(Precisazione del tema prescelto per la Tesi finale; inquadramento delle attività già svolte e da compiere nell'ultimo anno, con aggiornamento delle previsioni su obiettivi e metodologia; cronoprogramma).*

Applications of AI and ML will continue to transform the corporate sector as technology develops. For instance, because of its applicability across sectors, the use of AI and ML in forecasting is of enormous importance to most businesses. In the past, businesses depended on statistical forecasting techniques like linear regressions and exponential smoothing to aid in decision-making. But across companies and sectors, machine learning-based forecasting has largely taken the role of conventional methodologies in data and analytics initiatives. Therefore, the chosen theme for the final thesis centres on advancing machine learning in the transportation sector, with a specific focus on time series analysis, conformal prediction, and model interpretability. In the preceding phases of my PhD, I have engaged in several research activities. Initially, I developed a spatio-temporal demand modelling framework. Here, by leveraging deep learning models, I delved into how the granularity of space and time influences the prediction accuracy. The results, particularly those pertaining to the bidirectional LSTM's capabilities in capturing temporal patterns, were revealing. In another study, the use of ML in transport mode detection were explored. Moving beyond merely detecting the transportation mode, we proposed an innovative method that counts the number of metro stations in a metro trip using the data from mobile magnetometer sensors using the combination of an unsupervised machine learning method together with a post-processing algorithm. The method showcased promising accuracy proving to be an alternative method for counting the metro stations where GPS or Wi-Fi positioning are unavailable. Lastly, my focus shifted to road traffic accidents, a pressing global issue. Using a comprehensive dataset, we combined innovative techniques like one-hot encoding, SMOTE, conformal prediction, and compared various state of the art ML classification algorithms, road accident severity was predicted. The study's novelty was amplified by employing SHAP for model interpretation, ensuring transparency and insight into influential factors. As I transition into the final year of my PhD, my research will be focused more into implementing conformal prediction for time series data in transportation. Several challenges, such as stochastic variations and the intricate spatial-temporal correlations among regions, are foreseen. However, leveraging advanced methodologies like ENBPI, SPCI, AdaptCI, and HopCPT, I am confident of navigating these challenges adeptly. Concurrently, I intend to undertake a comparative study, evaluating various uncertainty quantification methods. This will be accomplished by using two real-world datasets, shedding light on the intricacies of traffic flow, speed data, and taxi demand. Anomaly detection also forms a crucial part of my research trajectory. By detecting irregularities in transportation data, my goal is to enhance the accuracy of forecasting models and ultimately improve traffic safety. The pinnacle of my efforts will be the creation of an interpretable machine learning framework. This will be more than just a prediction tool; it will be a thorough model that provides clear insights, fosters transparent decision-making, and builds confidence among stakeholders.

**2 – Attività di ricerca realizzata nei primi due anni** *(identificazione e documentazione delle attività di: raccolta dati, sviluppo modelli, calibrazione, validazione delle procedure, eventuali criteri di autoverifica, etc.).*

Below is a combined procedure carried out in the first two years of the research:

**1. Data Collection:**

Datasets Used:

-Taxi demand data from New York City.

- Mobile magnetometer sensor data of metro riders in Sussex, Rome and Stockholm.

- Road traffic accident dataset from Rome (2006-2022).

**2. Model Development:**

Spatio-Temporal Demand Modelling Framework:

-Deployed deep learning models including Long Short-Term Memory (LSTM), Convolution Neural Networks (CNN), and Temporal-Guided Networks (TGNet).

- Implemented grid-based tessellation strategy for clustering the area.

Transport Mode Detection:

-Utilized mobile magnetometer sensor data.

-Extracted contextual features to recognize acceleration states.

-Implemented k-means unsupervised method for data classification.

-Developed station counter algorithm.

Traffic Accident Severity Prediction:

-Implemented machine learning models including the Extreme Gradient Boost (XGBoost), Decision Trees (DTs), Random Forest (RF), Extremely Randomized Trees (ETC) and Light Gradient Boosting (L-GBM).

-Introduced one-hot encoding for categorical variables.

-Utilized Synthetic Minority Over-Sampling Technique (SMOTE) to address data imbalance.

-Implemented conformal prediction to quantify prediction uncertainty.

-Employed SHapley Additive exPlanations (SHAP) for model interpretability.

**3. Calibration:**

Spatio-Temporal Demand Modelling:

-Adjusted model parameters to optimize prediction performance for different spacetime granularities.

Transport Mode Detection:

-Tuned algorithm parameters to enhance station counting accuracy.

Traffic Accident Severity Prediction:

-Calibrated model parameters for superior predictive accuracy and reduced uncertainty.

**4. Validation of Procedures:**

Demand Modelling:

-Evaluated model performance against actual taxi demand data.

-Assessed prediction accuracy for different space and time granularities.

Transport Mode Detection:

-Validated station counting algorithm against actual station numbers in Rome and Stockholm metro systems.

Traffic Accident Severity Prediction:

-Benchmarked model predictions against actual accident severity data.

-Assessed the influence of various factors on model prediction accuracy using SHAP.

**5. Self-Verification Criteria:**

-Internal Consistency Checks: Ensured consistent data processing and modelling steps across different datasets and models.

-Out-of-sample Validation: Used separate training and testing datasets to validate the performance of developed models.

-Resampling Techniques: Deployed techniques like cross-validation to ensure model robustness and reduce overfitting.

-Performance Metrics: Utilized metrics such as Mean Absolute Error, Root Mean Squared Error, and Accuracy to evaluate model performance and refine models accordingly.

**3 – Esame delle problematiche emerse e degli aspetti critici** *(breve discussione degli elementi caratterizzanti il lavoro compiuto, con particolare attenzione agli aspetti più critici ed alle difficoltà emerse, con indicazione delle soluzioni individuate o delle alternative praticabili per la prosecuzione delle attività).*

Spatio-Temporal Demand Modelling Challenges:

Granularity Issue: Determining the optimum spatial and temporal granularity for effective demand forecasting was challenging.

Solution: Extensive experiments were conducted using various combinations of space and time granularity, enabling us to identify optimal settings for different deep learning models.

Transport Mode Detection Difficulties:

Data Limitations: Unavailability of labelled data for validating the counting algorithm.

Solution: We created our own datasets for two different locations (Rome and Stockholm) by monitoring and labelling the magnetometer data by noting the time at which the metro stops and accelerates.

Traffic Accident Severity Prediction:

Data Imbalance: Accident severity datasets often have a skewed distribution, which can bias the predictive models. **Solution**: Implementation of the Synthetic Minority Over-sampling Technique (SMOTE) effectively addressed this imbalance, enhancing model performance.

Model Uncertainty: Traditional predictive models do not provide insights into the uncertainty of their predictions. **Solution**: Integrating conformal prediction effectively quantified prediction uncertainty, thereby enhancing decision-making reliability.

Black-Box Models: Deep learning models, while effective, often lack transparency, making it hard to interpret their predictions. **Solution**: The application of SHapley Additive exPlanations (SHAP) ensured model interpretability, providing insights into factors influencing predictions.

**4 – Potenzialità di conseguire un "impatto" scientifico significativo** *(giudizio critico sulla efficacia ed originalità che la ricerca, al termine del Dottorato, potrà dispiegare, in relazione al quadro scientifico di riferimento e all'evoluzione delle conoscenze in corso in ambito nazionale ed internazionale).*

The research undertaken during this doctoral journey holds the potential to achieve a significant scientific impact in multiple ways:

Addressing Fundamental Challenges: The research confronts fundamental and pressing challenges in the domain of transportation engineering and modeling. By examining and refining granularity in spatio-temporal modeling, there's an opportunity to set a new precedent in how transportation demand is forecasted. This is critical in a world that is increasingly urbanized and requires efficient transportation networks.

Innovative Application of Mobile Data: The work on using mobile magnetometer sensor data for detecting metro stations paves the way for a deeper integration of mobile sensors into transportation analytics. This can revolutionize underground transportation analysis, especially in environments where traditional GPS or Wi-Fi-based systems fail.

Enhanced Safety Measures: The work on predicting road traffic accident severity, combined with the application of one-hot encoding, SMOTE, and conformal prediction, introduces a robust framework for urban safety. Given the global concerns around road traffic accidents, this research has the potential to be adopted and adapted by cities globally, contributing to saving lives and reducing injuries.

Interdisciplinary Integration: The integration of advanced machine learning, deep learning, and conformal prediction methodologies demonstrates the effectiveness of interdisciplinary research. This convergence is likely to inspire further integration of machine learning and traditional engineering domains, leading to more robust and accurate models.

Contribution to Conformal Prediction: With the emphasis on implementing conformal prediction for time series data, this research can significantly contribute to the literature on conformal prediction, especially in the context of transportation. The solutions identified to address challenges like stochastic variations and spatial-temporal correlations can be referenced and built upon in future studies.

Global Relevance: The challenges addressed by this research are not unique to any one region or country. Urbanization, the need for effective transportation, and concerns about road safety are global. Thus, the findings and methodologies of this research have the potential to be applied internationally, further amplifying its impact.

Setting a Foundation for Future Research: The methodologies developed and refined during this doctoral study, particularly the machine learning framework, offer a foundation for future researchers to build upon. The comparative studies and anomaly detection methods proposed as future steps will further enhance this foundational contribution.

In light of the evolving knowledge in transportation engineering and machine learning at both national and international levels, this research exhibits originality, depth, and applicability. It's positioned to not just contribute to the academic discourse but also to have practical implications for policymakers, city planners, and transportation professionals globally. In essence, the research undertaken during this doctorate has the potential to shape the future trajectory of transportation modeling and safety analysis.

## 5 – Schema di impostazione della Tesi finale di Dottorato e programmazione delle attività di completamento.

**1. Title Page:**

Title of the thesis

Name of the candidate

Affiliation (university/institute)

Date of submission

**2. Acknowledgments:**

Expressing gratitude to advisors, colleagues, funding sources, and any others who supported the research.

**3. Abstract:**

Brief summary of the research, methodologies, key findings, and contributions.

**4. Table of Contents:**

Organized list of chapters, sub-sections, and appendices.

**5. Introduction:**

Background of the research domain.

Statement of the research problem.

Research objectives and significance.

Brief overview of the research methodology.

Outline of the thesis structure.

**6. Literature Review:**

Comprehensive review of existing studies related to the research topic.

Identification of gaps in the current literature.

Justification for the current study.

**7. Methodology:**

Detailed description of data collection procedures.

Explanation of model development, calibration, and validation.

Elaboration on tools, software, and statistical tests used.

**8. Data Analysis and Results:**

Presentation of results from each study conducted.

Comparative analyses, where relevant.

Visual aids like graphs, charts, and tables to illustrate findings.

**9. Discussion:**

Interpretation of the results in light of the research questions.

Comparison with findings from the literature.

Examination of implications and potential applications of the research.

**10. Challenges and Critical Aspects:**

Discussion on the difficulties encountered during the research.

Solutions and alternatives employed to address these challenges.

**11. Potential Scientific Impact:**

Reflection on the expected contribution of the research to the field.

Discussion on the broader implications of the findings.

**12. Conclusion and Recommendations:**

Summary of the main findings.

Recommendations for policymakers, practitioners, and future researchers.

Suggestions for potential areas of future research.

**13. Future Work:**

Detailed plans for subsequent research or applications based on the current findings.

**14. References:**

List of all sources cited in the thesis.

**15. Appendices:**

Supplementary material, such as raw data, additional analyses, or code snippets.

**6 – Cronoprogramma** *(seguire lo schema seguente)*

| **n.** | **Attività** | **II Anno (consuntivo)** | | | | **III Anno** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **I** | **II** | **III** | **IV** | **I** | **II** | **III** | **IV** | | | | |
| **1.** | Internship at PTV | **X** | **X** | | | | | | | | | | |
| **2.** | Literature review | **X** | **X** | **X** | | | | | | | | | |
| **3.** | Data collection and analysis | **X** | **X** | **X** | | | | | | | | | |
| **4.** | Model Development: | **X** | **X** | **X** | | | | | | | | | |
| **5.** | Model Calibration: | **X** | **X** | **X** | | | | | | | | | |
| **7.** | Transport Mode Detection | **X** | | | | | | | | | | | |
| **8.** | Road Traffic Severity Prediction | | **X** | **X** | | | | | | | | | |
| **9.** | Papers Submission | **X** | | **X** | | | | | | | | | |
| **10.** | Initiate Conformal Prediction for Time Series | | | | **X** | | | | | | | | |
| **11.** | Data Analysis for Time Series | | | | **X** | | | | | | | | |
| **12.** | Comparative Study | | | | **X** | **X** | | | | | | | |
| **13.** | Anomaly Detection | | | | | **X** | **X** | | | | | | |
| **15.** | Interpretable Machine Learning Framework | | | | | | **X** | **X** | | | | | |
| **16.** | Thesis Compilation | | | | | | **X** | **X** | | | | | |
| **17.** | Final Reviews and Revisions | | | | | | | **X** | **X** | | | | |

## SEZIONE B

## Attività di collaborazione e supporto; formazione ed acquisizione di capacità evolute

**(massimo 2 pagine)**

**1 – Partecipazione alle attività di didattica presso la struttura di afferenza** *(attività seminariale, supporto alla didattica frontale, preparazione di materiale didattico, collaborazione per ricevimento studenti, collaborazione allo svolgimento di tesi di laurea e stages).*

- Assistance in Transportation and Modelling exam for Master students conducted on July 2023 and September 2023.

- Supervising a master students during their thesis work:

  - "Theoretical and Practical Analysis of PTV Balance A Well Proven Network-Wide Solution for Optimization and Coordination of Signalized Intersections" By graduate student Ramin Bohlouli

**2 – Attività di formazione** *(soggiorni presso strutture di didattica e ricerca in Italia e all'estero, corsi curriculari o speciali frequentati, partecipazione a seminari, convegni, workshop, etc.).*

**A. Courses:**

-Artificial Intelligence A-Z 2023 by Udemy

- Html, CSS and JavaScript

- PTV Academic Exercises – PTV Vissim

- PTV Academic Exercises – PTV Visum

- PTV Academic Exercises – PTV Viswalk

**B. Conferences:**

-2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)

**C. Internship:**

Collaboration with PTV SISTeMA:

-Assisted in understanding and maintaining the software by identifying and resolving software bugs.

-Provided exceptional support to clients, addressing their inquiries and concerns.

-Developed and implemented robust Continuous Integration/Continuous Deployment (CI/CD) pipelines using Jenkins for Balance and Epics software.

**D. Seminars:**

-Public Transportation: (Low) fares, equity and policy- Workshop at the 11th Symposium of the European Association for Research in Transportation at ETH Zurich.

- Multi-scale modelling of active mode traffic and transportation: no data, no glory! - Prof. Serge Hoogendoorn – TU Delft

- Control and data in ITS: from Big Data to Smart Data – Prof. Arnaud de la Fortelle – Mines Paris PSL & Heex Technologies

- Traffic Flow of Urban Air Mobility: Modeling, Control, and Simulation - Prof. Jack Haddad - Technion University

**3 – Collaborazione a studi, ricerche, programmi strutturati** *(contributi in PRIN, ricerche di Facoltà e di Ateneo, convenzioni, etc., con inquadramento del programma e specificazione dell'attività prestata).*

## SEZIONE C

## Informazioni

*(Tale sezione contiene le informazioni richieste alla fine ogni anno dall'Ufficio Dottorati)*

1) Titolare di borsa erogata dalla Sapienza - Università di Roma……………SI☑ NO□

2) Nazionalità ……Indian……………………………………………………..

3) Dottorato in cotutela ………………………………………….……………SI□ NO☑

(se si indicare il cotutore…………………………..)

4) Dottorato con doppio titolo ……..…………………………….……………SI□ NO☑

5) Borsa con finanziamento esterno ……………………………….……………SI□ NO☑

6) Università di provenienza ………"La Sapienza" University of Rome

7) Numero di mensilità di ricerca spese in una struttura di ricerca estera ……0…

8) Finanziamenti all'interno di reti internazionali di formazione alla ricerca ..SI□ NO☑

9) Pubblicazioni e altri prodotti degli ultimi 3 anni

***Per le aree bibliometriche****. Articoli pubblicati su riviste peer-reviewed internazionali (ed eventualmente proceedings per le aree che accettano) con impact factor (indicizzate WoS) o indicizzate Scopus.*

-Varghese, Ken Koshy, Mahdaviabbasabad, Sajjad, Gentile, Guido and Eldafrawi, Mohamed. *"Effect of Spatio-Temporal Granularity on Demand Prediction for Deep Learning Models"* Transport and Telecommunication Journal, vol.24, no.1, 2023, pp.22-32. https://doi.org/10.2478/ttj-2023-0003.

- S. H. Hosseini, G. Gentile, K. K. Varghese and L. M. B. Miristice, *"Inferring Station Numbers in Metro Trips Using Mobile Magnetometer Sensor via an Unsupervised K-means Clustering Algorithm,"* 2023 8th

International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Nice, France, 2023, pp. 1-6, doi: 10.1109/MT-ITS56129.2023.10241558.

- Mohamed Eldafrawi, Ken Koshy Varghese, Marzieh Afsari, Mahnaz Babapourdijojin, Guido Gentile *“Predictive Analytics for Road Traffic Accidents: Exploring Severity through Conformal Prediction.”*
Conference: Accepted for 2024 TRB Annual Meeting.

***Per le aree non bibliometriche**. Prodotti editoriali pubblicati dai dottorandi come Monografie dotate di ISBN e/o pubblicazioni in riviste di fascia A (o prodotti editoriali equivalenti ammessi dalla VQR).*