Executive Summary

Urban transport planning is inherently complex, requiring accurate representation of transport demand and mobility patterns across spatial and temporal dimensions. Intelligent Transport Systems (ITS) have emerged as pivotal in enhancing the efficiency and environmental sustainability of urban transport networks by automating data collection and processing. Modern technologies, such as mobile applications, Geographic Information Systems (GIS), Internet of Things (IoT) devices, and 5G connectivity, have significantly improved data quality and quantity, outperforming traditional methods like telephone surveys. The transport sector has used data collection methods such as passenger counting (APC) systems to estimate mobility patterns and improve public transport services. Numerous studies in literature have sought to refine the data collected through these systems, aiming to provide a more precise understanding of mobility trends.

While existing literature has focused on enhancing univariate data for prediction and categorisation, efforts have also been made to integrate external variables like environmental and meteorological data to improve prediction accuracy. However, despite these advancements, selecting appropriate systems or processes remains challenging due to the lack of standardised protocols, often resulting in sub-optimal implementations. Moreover, determining the best method for specific datasets is still difficult, as few studies have compared these models using real operational data or considered factors like dataset size and computational power.

This research addresses these gaps by empirically evaluating passenger count estimation models using both artificial intelligence (AI) and traditional statistical techniques, with a focus on data fusion, using real operation data. It also examines the discrepancies between claimed and actual performance of Automated Passenger Counting (APC) systems, which often report accuracy rates of 98% to 99%, yet real-world performance varies widely between 50% and 78%. Specific objectives include developing and implementing diverse passenger count estimation models, applying methods from other fields to the transport sector, evaluating model performance through comprehensive metrics, and providing a comparative analysis to identify each model's strengths and weaknesses in passenger count estimation.

An eight-step methodology was employed to achieve these objectives, involving data collection, processing, and the application of multiple data fusion models. The study tested different models using a data fusion approach that integrated multiple data sources, including two passenger count systems (one camera-based and one Wi-Fi-based), Points of Interest (POIs), meteorological data, pollution data and 57 days of ground truth data collected manually by the author during operation times in Turin. Statistical models were compared against machine learning

and deep learning models, with results evaluated using a decision matrix approach based on various factors and potential transport companies' requirements, to then rank the different models.

The research found that LightGBM was the best-performing model, followed by Bayesian regression and XGBoost, based on accuracy, error and scalability metrics across different experiments. Bayesian regression showed strong performance even with smaller datasets, while SARIMAX and Prophet models benefited from data augmentation. Advanced time series models like LSTM and Prophet were less effective, underscoring the importance of aligning algorithm choice with dataset characteristics such as size and seasonality.

The study demonstrates that high-quality, well-distributed data is crucial, and ensemble machine learning algorithms hold significant promise for data fusion in the transport sector. The findings suggest that further research should focus on refining data collection and cleaning methodologies to enhance model performance and exploring time-series model for vehicle load estimation. The research contributes to both the academic field and practical applications by providing insights into reality representation and offering a framework for transport companies to select suitable models for passenger load estimation.

In conclusion, this research provides a thorough evaluation of various data fusion models and offers practical guidelines for their application in the transport sector. The results underscore the importance of data quality and strategic algorithm selection to optimise transport system performance. Further studies should continue to explore advanced data analytics and to select the most effective (typology and size) data sets to address the evolving challenges of urban mobility and to evaluate the reality representation of these models in practical implementations within the field.