TOWARDS PREDICTIVE SITUATIONAL AWARENESS OF URBAN PEDESTRIAN FLOWS: AN ASSESSMENT OF DATA QUALITY ON PERFORMANCE METRICS OF PREDICTIVE MODELS

Carrow Morris-Wiltshire[1]
✉ c.morris-wiltshire@ncl.ac.uk

Professor Stuart Barr[1], **Professor Phil James**[1], **Keith Hermiston**[2]

[1] School of Engineering, Newcastle University, Newcastle upon Tyne, UK
[2] DSTL, Newcastle upon Tyne, UK

Introduction:

Urban observatories (UOs) are emerging as data powerhouses in many UK cities, marking a transformative era in how we understand and manage urban life. These UOs are complex systems that continuously gather a vast array of data from air quality to traffic patterns. However, tapping into this treasure trove of data is not straightforward; it requires specialised skills and resources. Moreover, for this data to benefit the widest possible audience, it needs to be easily accessible and usable—even for those without expert knowledge.

One key challenge is the accuracy of the collected data, often compromised by faulty sensors (James, Jonczyk et al. 2022). This thesis explores how artificial intelligence can step in to automatically identify and label such erroneous data, making UO datasets more reliable and user-friendly.

Research Objectives:

1. Develop a modelling workflow that identifies unusual patterns (anomalies) in the data using single-step univariate prediction validating on unseen data.
2. Generate additional input features (feature engineering) to assess how the model's performance changes for anomaly prediction.
3. Measure the change in prediction accuracy as the prediction horizon increases for univariate and multivariate models.
4. Measure the change in prediction accuracy as the data completeness for the training data is reduced for univariate and multivariate models.

Methodology:

The workflow that has been developed compares a baseline linear model with a simple LSTM (long short-term memory) with both configured for univariate and multivariate modelling. 1310 models were trained using a range of hyperparameters to answer the research questions. Each model was then evaluated over 5 runs to account for the model stochasticity.

Results:

- Using multivariate inputs makes the models more robust for predictions on lower quality data.
- The best performing models used 30 days of training data with 98% daily completeness.
- The test models can accurately make predictions up to an hour and a half into the future on data with 50% completeness.

Conclusion:

The findings reveal that a small amount of high-quality data suffices to build a reliable predictive model for pedestrian behavior from urban observatory sensors. Features that capture natural cycles in the data enable the model to accurately identify anomalies, even in incomplete datasets.

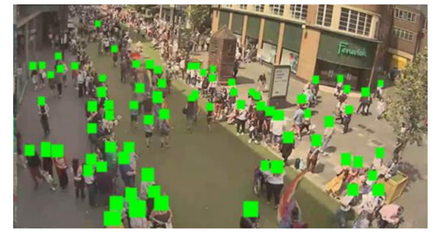
By considering **natural rhythms** in the data, multivariate **predictive models** can successfully **identify** unusual **patterns** in **incomplete** pedestrian **data** from urban observatory sensors.

EDA Report

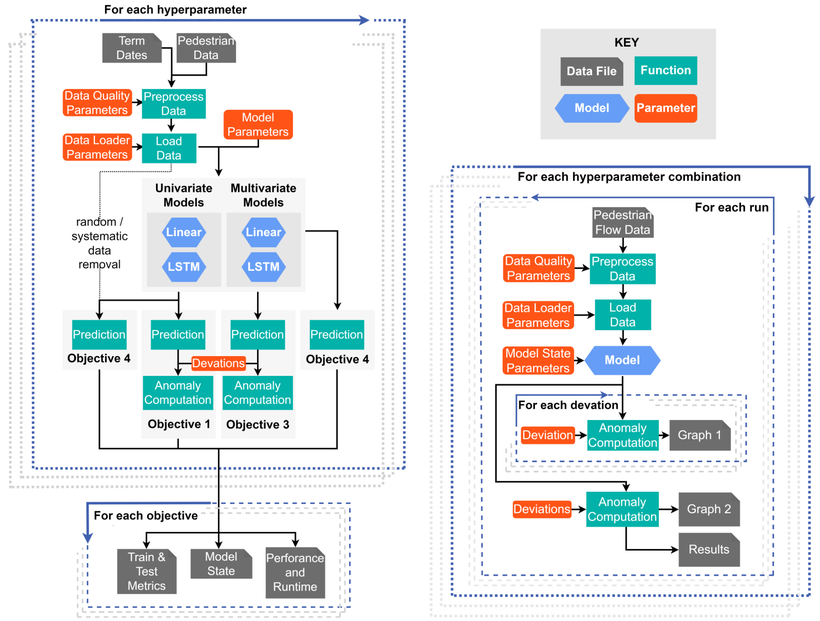
Thesis

GitHub Repo

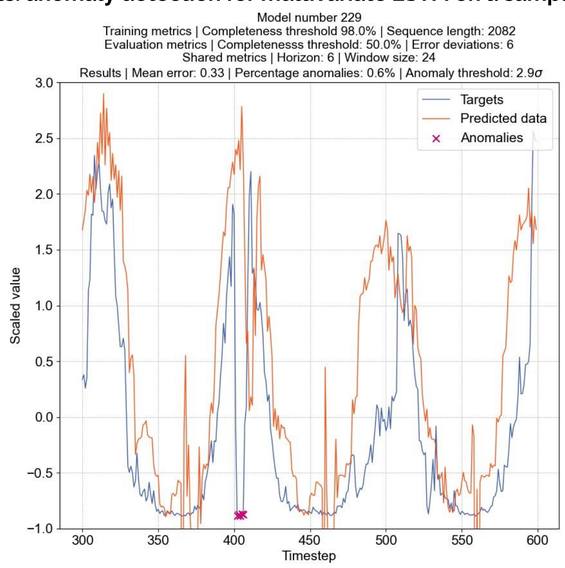
Background: pedestrian counting using object detection in action



Methodology: workflow for training (left) and evaluation (right)



Results: anomaly detection for multivariate LSTM on a sample sequence



References

James, P., et al. (2022). "Realizing Smart City Infrastructure at Scale, in the Wild: A Case Study." Frontiers in Sustainable Cities **4**: 767942.

[dstl] GEOSPATIAL SYSTEMS Newcastle University urban observatory