Annual Progression Report 1

Produced by

Carrow Morris-Wiltshire

Supervised by *Prof. Stuart Barr* and *Prof. Phillip James*

Sponsored by DSTL

June 2024

Newcastle University

1 Introduction

Predicting short-term behaviour of agents within a city requires reliable high-quality data, real-time processing infrastructure, and models that are capable of generating accurate predictions about complex situations at speed. Enabling this technology could have profound benefits for cities in domains such as emergency response, resource optimisation and decision-making, thus allowing urban areas to better to respond to unsafe situations such as overcrowding or flooding.

Centralised data repositories are appearing in cities around the world, collecting near real-time data from distributed sensor networks. However, issues with data quality need to be addressed before predictive systems can be developed. The volume and velocity of the data collected by these sensors requires automated quality-aware systems to assess veracity before the data can be incorporated into the decision-making systems enabled by short-term predictive capabilities.

Whilst there is existing work in this area but it has not yet been applied at scale to data collected by urban sensor networks.

2 Aim, Objectives and Scope

2.1 Aim

To develop an AI system for the prediction of spatiotemporal dynamics in the built environment using near real-time geospatial IoT sensor data.

2.2 Research Objectives

- RO1: Develop tools to assess the quality of near real-time sensor data.
- RO2: Assess the spatiotemporal dependency of near real-time sensor data.
- RO3: Assimilate outputs of agent-based models with real-time sensor data to monitor urban systems in real-time.
- RO4: Evaluate the approach using real-world case-studies and develop a roadmap for scalable deployment.

2.3 Scope

By applying existing data-quality frameworks to develop a data quality monitoring system RO1 seeks to quantify data quality in a manner that can both be interpreted by an end-user (dashboards) and a programme interface for use in later objectives. Making predictions requires an understanding of the dependency of the sensors in time and space – how strongly do sensor measurements relate to the wider sensor network – a spatial AI model in combination with existing knowledge about agent behaviour will be investigated (RO2). To enhance prediction, a data assimilation approach using the outputs of much larger computational models (such as an agent-based model of transport demand) will be investigated (RO3). Ensuring this research has real-world applicability and elevating the technology to a level that meets the needs of its users, a roadmap will be developed. A variety of computational models and sensor network combinations will be investigated to achieve this (RO4).

2.4 Deliverables

The research aims to contribute the following to the field of complex systems modelling:

- Greater understanding about the real dynamics of complex urban systems.
- An enhanced understanding of causality in complex urban systems.

And to deliver the following technical capabilities:

- A demonstration of value for the data collected by centralised urban repositories.
- Pathways to making this data 'AI ready'.
- A demonstration of a real-time cloud-based web-app that provides useful information derived from the sensor data that can be used for improved decision making.
- A package of code that conforms to best-practice and is built to be compatible with urban digital-twin frameworks such as DAFNI/Gemini.
- A series of publications showcasing any scientific advancements made by this research.

3 Approach and Methodology

3.1 Overview

The approach to this project remains aligned with the original proposal. The project is divided into four work packets based on the four research objectives. A brief overview of the proposed approach to each objective is provided below.

NOTE: All objectives will broadly follow MLOps best practices, including version control, continuous integration, and automated testing. The project will be developed using Python, with a focus on open-source libraries and platforms like PyTorch and Kubernetes. An outline of the pipeline architecture that will serve as the backbone for this project is provided in 1.

3.2 Objective 1: Tools for Data Quality Assessment

Developing tools that assess the quality of near real-time sensor data requires understanding the specific metrics and characteristics that define data quality. These metrics

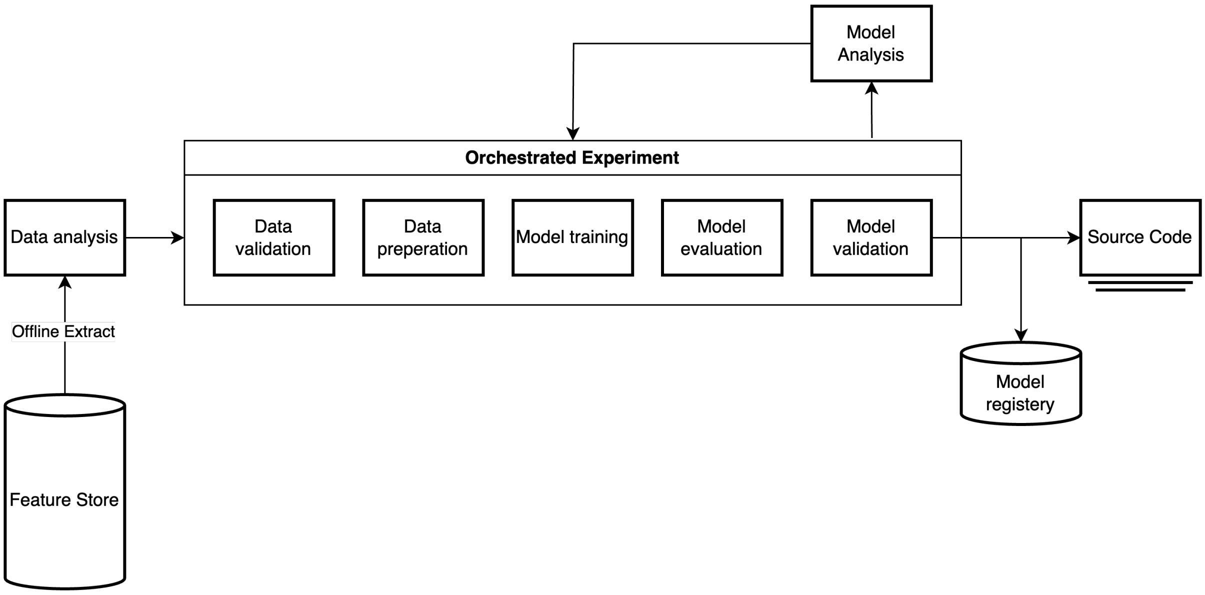


Figure 1: Proposed MLOps pipeline architecture for the project.

typically include accuracy, precision, timeliness, completeness, and reliability Wang & Strong (1996). The approach involves defining these quality parameters clearly, identifying the appropriate sensors and data sources, and then developing algorithms and tools to evaluate these aspects effectively.

The proposed approach begins with practical application case studies, such as pedestrian sensor data in railway stations to detect overcrowding. Data collection will use automated extraction tools for metrics like completeness and manual assessments for sensor-specific attributes such as object detection accuracy. To quantify DQ metrics like timeliness and accuracy, models such as those proposed by Fizza et al. (2022), who compute age of data to manage uncertainty during decision-making processes, will be used.

Comparative model evaluation will assess the impact of DQ metrics on prediction accuracy using statistical models like SARIMA and machine learning models including LSTMs and transformers. Metrics such as RMSE and R squared will be employed to evaluate model performance.

The methodology emphasises iterative testing and validation, ensuring that developed tools are robust and reliable. Initial findings will inform refinements, enhancing the tools' applicability across different IoT contexts, thereby ensuring comprehensive and accurate assessments of sensor data quality.

3.3 Objective 2: Spatio-Temporal Dependency Analysis

This objective focuses on understanding the spatiotemporal dependencies within IoT data streams to determine sensor coverage adequacy. The methodology starts with an exploratory analysis to estimate expected data pattern lags, such as walking or driving times between sensors. This preliminary analysis sets the stage for training a deep learning model, like a graph neural network (GNN), to make spatiotemporal predictions.

An example case study involves a network of pedestrian sensors in a city centre. By analysing data counts and movement patterns, the network's spatial distribution and reliability can be evaluated. The primary challenge is limited sensor distribution, which necessitates innovative pre-processing techniques. Methods like principal component analysis and Fourier analysis will decompose temporal signals, improving model performance (Kutz et al. 2016).

To assess spatiotemporal components, statistical methods such as spatial autocorrelation and dynamic time-warping Froese et al. (2020), alongside deep-learning approaches like hybrid GNNs, will be employed. Given the non-Euclidean nature of pedestrian networks, conventional CNNs are unsuitable (Klemmer et al. 2019). Instead, GNNs and GANs with local autocorrelation, as indicated by recent research, will be explored.

The model's ability to predict spatiotemporal dependencies will be validated using performance metrics like RMSE and MAE. The goal is to establish a robust framework for analysing spatiotemporal dependencies, providing insights into the adequacy of sensor networks and informing future IoT deployments.

3.4 Objective 3: ABM Assimilation

The methodology for integrating agent-based model (ABM) outputs with real-time sensor data involves several key steps. First, a comprehensive review of existing surrogate modelling techniques will identify potential enhancements. The primary approach involves training deep learning models on ABM outputs, focusing on architectures capable of capturing high-dimensional, nonlinear relationships in urban dynamics.

Various deep learning architectures, such as hybrid GNN-LSTMs and GANs with local autocorrelation, will be explored. The ABM will be run offline to produce a spatiotemporally rich training set, which will train the surrogate model (Kieu et al. 2022). This surrogate will then use real-time IoT sensor data to make predictions. For instance, pedestrian sensor data will be used to evaluate the surrogate's ability to predict pedestrian counts.

The model's performance will be evaluated on its ability to replicate ABM behaviours and make real-time predictions. The network of sensors will be split into training, testing, and validation sets. The surrogate's predictions will be validated using metrics like RMSE, MAE, and R squared. The model's ability to generalise across different sensor types and urban environments will also be assessed.

This methodology aims to create a computational bridge between ABMs and real-time IoT data, enabling efficient and accurate simulations of urban dynamics (Heppenstall et al. 2021). The end goal is a robust surrogate model capable of real-time predictions, enhancing urban monitoring and decision-making processes.

3.5 Objective 4: Approach Evaluation and Deployment Roadmap

The final objective focuses on evaluating and scaling the developed methodologies using real-world case studies. The methodology involves collaboration with academic partners and stakeholders, such as Newcastle City Council and national data science institutes like DAFNI and the Alan Turing Institute. Initial stakeholder engagement will identify critical use cases, guiding the research focus.

A key aspect of this objective is a placement in a city with advanced IoT infrastructure, such as Singapore or Melbourne, providing a real-world testbed for the developed tools. The research will involve reflective evaluation to understand the scalability and generalisability of the methodologies. This includes testing the surrogate models on different ABMs and sensor networks, assessing performance across varying spatial and temporal resolutions.

Developing a roadmap for scalable deployment involves creating a translation framework to address challenges such as industry standards compliance and risk mitigation. This framework will demonstrate integration with existing digital infrastructure like DAFNI and explore improvements in computational efficiency, potentially through code optimisation.

Engagement with stakeholders will ensure the research addresses practical needs, facilitating the transition from pilot projects to deployable solutions. The objective aims to produce a pilot software library for real-time situational awareness and a roadmap for achieving higher technology readiness levels, paving the way for broader application and impact.

4 Progress to Date

4.1 Overview

- A first draft of the literature review has been completed for the first research objective.
- The codebase has been largely built with regular commits on GitHub.
- A dashboard has been developed to demonstrate some of the functionality of the data quality measurement system.
- My research was published and presented at the 32nd GISRUK Conference 2024.

4.2 CPD

Up-skilling has been a major focus since January. Building scalable infrastructure requires knowledge of programming best practices beyond what was taught during the MRes. This includes understanding functional modularity, config-driven development, and object oriented programming paradigms, as well as specialised machine learning knowledge, for example, building Graph Neural Networks. I have completed a number of online, book-based and in-person courses to develop my programming and presentation skills, as well as attending all CDT organised events.

External courses and training:

- Hands-On Graph Neural Networks Using Python

- Introduction to SQL
- Hierarchical and Recursive Queries in SQL Server
- Object-Oriented Programming in Python
- Introduction to LLMs in Python
- Introduction to Testing in Python
- Building Dashboards with Dash and Plotly
- Python Toolbox
- Monitoring Machine Learning Concepts

Newcastle University courses and workshops:

- Time Series Data - MAS8382
- The Introduction to Learning and Teaching (ILTHE)
- Building an Impactful Presentation: A Step-by-Step Guide
- Delivery Skills: Master the Art of Effective Presentations
- Storytelling for Researchers: Unleash the Power of Narrative
- Designing an Effective Research Poster

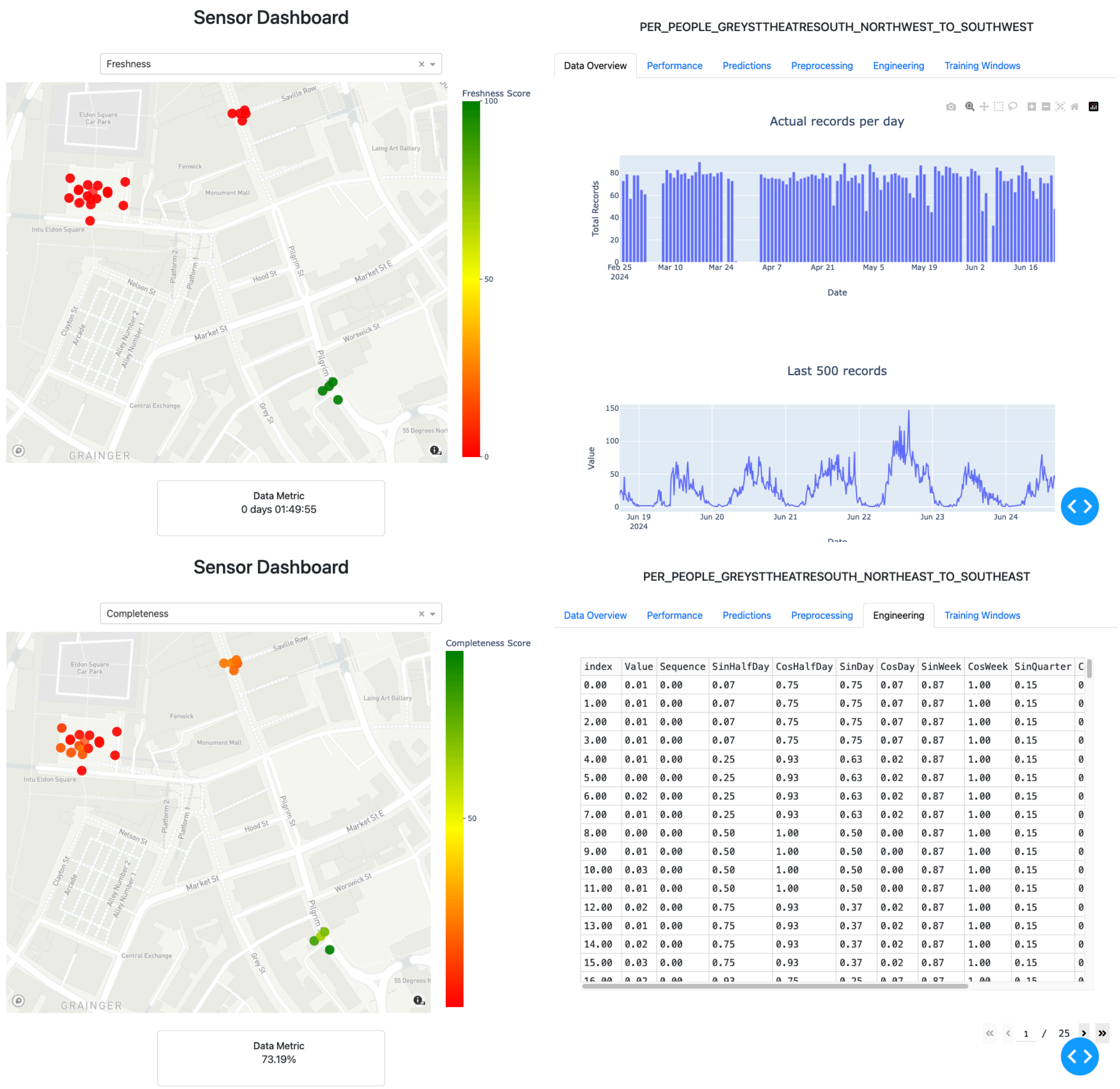
4.3 Codebase and Dashboard

The codebase which is primarily available here has been built with regular commits on GitHub. The codebase is structured to include the following components:

Figure 2: Top-level directory structure of the codebase.

```
phd/
├── apps/
├── configs/
├── data/
├── logs/
├── scripts/
├── src/
│   ├── api/
│   ├── data_processing/
│   ├── models/
│   ├── pipeline.py
│   ├── training/
│   ├── utils/
│   └── visualisation/
└── tests/
```

The dashboard is built using Dash and Plotly:



index	Value	Sequence	SinHalfDay	CosHalfDay	SinDay	CosDay	SinWeek	CosWeek	SinQuarter	C
0.00	0.01	0.00	0.07	0.75	0.75	0.07	0.87	1.00	0.15	0
1.00	0.01	0.00	0.07	0.75	0.75	0.07	0.87	1.00	0.15	0
2.00	0.01	0.00	0.07	0.75	0.75	0.07	0.87	1.00	0.15	0
3.00	0.01	0.00	0.07	0.75	0.75	0.07	0.87	1.00	0.15	0
4.00	0.01	0.00	0.25	0.93	0.63	0.02	0.87	1.00	0.15	0
5.00	0.00	0.00	0.25	0.93	0.63	0.02	0.87	1.00	0.15	0
6.00	0.02	0.00	0.25	0.93	0.63	0.02	0.87	1.00	0.15	0
7.00	0.01	0.00	0.25	0.93	0.63	0.02	0.87	1.00	0.15	0
8.00	0.00	0.00	0.50	1.00	0.50	0.00	0.87	1.00	0.15	0
9.00	0.01	0.00	0.50	1.00	0.50	0.00	0.87	1.00	0.15	0
10.00	0.03	0.00	0.50	1.00	0.50	0.00	0.87	1.00	0.15	0
11.00	0.01	0.00	0.50	1.00	0.50	0.00	0.87	1.00	0.15	0
12.00	0.02	0.00	0.75	0.93	0.37	0.02	0.87	1.00	0.15	0
13.00	0.01	0.00	0.75	0.93	0.37	0.02	0.87	1.00	0.15	0
14.00	0.02	0.00	0.75	0.93	0.37	0.02	0.87	1.00	0.15	0
15.00	0.03	0.00	0.75	0.93	0.37	0.02	0.87	1.00	0.15	0
16.00	0.02	0.00	0.93	0.75	0.25	0.07	0.87	1.00	0.15	0

Figure 3: Dashboard

4.4 Literature Review

The first draft of my literature is available here along with this document. As part of a skills development exercise, I have been using GitHub to host and manage the literature review (and other LaTeX documents), developing a pdf viewer using JavaScript, and building in commenting functionality using hypothesis.is to enable concurrent feedback from both supervisors. The literature review is currently undergoing feedback from my supervisors and will be revised accordingly. The literature review is focused primarily on quality-aware data streams specifically in the context of wireless sensor networks for smart cities. The case-study that is being used in this research is the pedestrian counting system developed by the Urban Observatory and Newcastle upon Tyne. The literature review has been structured to cover the following areas:

1. *Data Quality Dimensions and Metrics*
2. *Data Quality Assessment*
3. *Data Quality Detection and Monitoring*
4. *Data Quality Management and Improvement*
5. *Data Quality Prediction and Proactive Approaches*
6. *Challenges and Future Directions*

5 Work Plan

My work plan is available as a web-app hosted on Google App Engine (allow up to 10s for the chart to load in). The slippage dropdown show the current estimates in red overlayed on the original estimates. There is also view to show the original timeline and the updated one separately. There is a possibility of a 6-month interruption starting in November that can be toggled on and off. The app was developed as a side project in Python to reinforce my knowledge from my Plotly-Dash course. I have designed the app as a Python package, available on my GitHub intended for use by other PhD students—I found there was a shortage of free and easy-to-use Gantt-(style) charts. This project helped me to understand how to build a Python package following best-practice and I intend to publish it on my PyPi in the near future.

6 Financial Expenditure

Category	**Balance**
CDT Events	*-£3,500.00*
Books/Period	*-£32.39*
Conferences	*-£185.00*
Computers	*-£2,984.91*
Subsistence	*-£139.47*
Other	*-£7.50*
Travel	*-£650.05*
Misc	*-£173.58*
Total RTSG	**£22,000.00**
Expenditure to date	**-£7,672.90**
Remaining top-slice	**-£3,500.00**
Balance Remaining	**£10,827.10**

Table 1: Expense Summary

7 References

References

Fizza, K., Jayaraman, P. P., Banerjee, A., Georgakopoulos, D. & Ranjan, R. (2022), Age of Data Aware Internet of Things Applications, *in* '2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)', pp. 399–404.

Froese, V., Jain, B., Niedermeier, R. & Renken, M. (2020), 'Comparing temporal graphs using dynamic time warping', *Social Network Analysis and Mining* **10**, 1–16.

Heppenstall, A., Crooks, A., Malleson, N., Manley, E., Ge, J. & Batty, M. (2021), 'Future Developments in Geographical Agent-Based Models: Challenges and Opportunities', *Geogr Anal* **53**(1), 76–91.

Kieu, M., Nguyen, H., Ward, J. A. & Malleson, N. (2022), 'Towards real-time predictions using emulators of agent-based models', *Journal of Simulation* pp. 1–18.

Klemmer, K., Koshiyama, A. & Flennerhag, S. (2019), 'Augmenting correlation structures in spatial data using deep generative models', *arXiv preprint arXiv:1905.09796* .

Kutz, J. N., Brunton, S. L., Brunton, B. W. & Proctor, J. L. (2016), *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*, SIAM.

Wang, R. Y. & Strong, D. M. (1996), 'Beyond accuracy: What data quality means to data consumers', *Journal of management information systems* **12**(4), 5–33.