Characterizing Multi-Source Data for Effective Urban Mobility Modelling: The Case of New York City

Tonny Rutayisire[1,*], Jun Liu[1,†], Samuel Moore[1,†] and Carlos Balsas[2,†]

[1]School of Computing, Ulster University, UK
[2]School of Architecture & Built environment, Ulster University, UK

Abstract

To explore a more holistic view of urban human mobility, a few researchers have previously attempted to build models driven by a fusion of multi-source datasets. Yet, owing to limited understanding of the intrinsic characteristics and relationships underlying multi-source datasets, most of such interventions naively merge these datasets, producing sub-optimal prediction models. To address this issue, we propose a three-step methodological framework to capture urban mobility from an integrative point of view. The novelty of our framework is three-fold: (i) a systematic characterization of the multi-source data to leverage hidden relationships when integrating the data; (ii) integrating data with contextual information of the urban environments to improve explainability; and (iii) conforming the model building to incremental learning paradigm to adapt to changing patterns of mobility data. Using the New York City (NYC) case study, extensive analyses show salient relationships within datasets which could form a basis for optimal data fusion and subsequently improving the mobility modelling pipeline in line with the proposed three-step methodological framework.

Keywords

human mobility, data fusion, multi-view modelling,

1. Introduction

Human mobility patterns in urban settings are indicative of various phenomena such as travel demand, traffic congestion, resource allocation, CO_2 emission, and infectious disease spread. Therefore, understanding such patterns is crucial to applications such as transportation management, urban planning, resource optimization, and epidemiology among others. In recent years, the pervasive use of sensing technologies has produced an enormous amount of human mobility data that researchers have used to answer critical questions pertaining to when, why, and how people move from one place to the other [1]. A wide range of data-driven models and techniques have been proposed to capture urban human mobility, based on a variety of dataset: taxi trip records [2] [3], call detail records (CDRs) [4] [5], social media [6], [7]. Despite the large number of studies on urban human mobility, majority of existing techniques have typically been built on single-source empirical data in isolation from other mobility flows. Human mobility, especially in urban settings is multi-faceted, mainly due to diverse travel modes and different spatial/temporal scales. It inevitably introduces a bias, against uninvolved flows, when the capture of urban-scale mobility is single source data-driven. To explore a more holistic view of urban mobility dynamics, a few researchers have attempted to build models and techniques driven by a fusion of multi-source datasets. Zhang et al. [8] proposed **coMobile**, a multi-view learning framework based on integration of transit view and cellphone view. The approach outperforms 2 single-view models **WHERE** and **TRANSIT** by 51% and 58% in terms of Mean Average Percetage error (MAPE), respectively. Jiang et al. [9] also utilized taxi GPS trajectories, smart card transaction data of subway and bus from Beijing to model human mobility in space. It is reasonable to assume that the fusion of these mobility datasets addresses the biases to some degree and provides a representative view of a broad spectrum of the population. Yet, owing to limited understanding of the intrinsic characteristics and relationships of

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
*Corresponding author.
†These authors contributed equally.
rutayisire-t@ulster.ac.uk (T. Rutayisire); j.liu@ulster.ac.uk (J. Liu); s.moore2@ulster.ac.uk (S. Moore); c.balsas@ulster.ac.uk (C. Balsas)

these multi-source datasets, a considerable number of such interventions naively merge these datasets, which end up introducing hidden drawbacks to the data fusion and subsequently producing sub-optimal prediction models. For instance, in scenarios where space is a limitation, merging all multi-source datasets can become an issue, yet correlations within different datasest can be leveraged to use part of the dataset to infer the rest. Moreover, simplistic merging of datasets could actually worsen the sampling bias problem where datasets with low levels of representativeness contribute equally or more, to the data fusion. To this end, we propose a three-step methodological framework to capture urban mobility from an integrative point of view: (i) In the first step, we leverage intrinsic characteristics of multi-source data to learn, select, and integrate the most optimal vectors of the data for effective urban mobility modelling; (ii) we then integrate the modelling process with prior knowledge contexts of the domain to minimize the amount of data requirement, manage the uncertainties, and achieve model interpretability; (iii) and in the third step, we conform the modelling process to incremental learning paradigm where the model learns and updates by integrating new data streams as and when they become available. For multi-view learning, we argue that the effectiveness of the model can be enhanced when the above-mentioned aspects are incorporated together in the modeling process.
The rest of the paper is organized as follow: Section 2 summarizes the challenges and our motivation for this research. In Section 3, we introduce our generic methodological framework for integrative human mobility modeling. Section 4 introduces our case study area, the mobility datasets, and the approaches we have used to characterize the multi-source datasets. The results and their discussions are presented in Section 5, while the conclusion and future work are provided in Section 6.

2. Motivation

Urban mobility is multi-faceted in nature, mainly due to different modes of transport. As such, we have recently seen a growing shift from single source to multi-source data-based models for human mobility modelling in urban settings [9] , [10], [11], [12]. In reality, single data sources provide a limited perspective on human mobility. For instance, taxi trip data may not represent the people who cannot afford taxi and choose to use other means of transport such as bus, ride-sharing, and subway. On the other hand, data from bike-sharing apps is not as representative during the winter season, and or across unfavourable topography. Figure 1 shows the differences in trip counts for Yellow taxi & Green taxi across NYC taxi zones. Even with the same mode of transport (but owing to different urban contexts), we observe a huge divergence in the flow of mobility produced by yellow and green taxi, hitting as high as 380,000 trips in some zones. Hence, whereas fusion of multiple data sources may provide a more representative understanding of urban mobility patterns, simply adding together a series of features extracted from multi-source datasets lays a sub-optimal foundation for multi-view mobility modeling. It can't simply be assumed that all multi-source datasets contribute equally to the mobility flow aspect being learned. In [11] , authors highlight that the representativeness of each source largely depends on the demographics of the service users in relation to the demographics of the local population, and this dynamic is yet to be fully investigated. An effective approach to multi-view mobility modeling has to capture the true spatial-temporal context of the datasets, and their respective qoutas for the fusion. We argue that leveraging intrinsic characteristics of the datasets, coupled with existing domain knowledge could result in an appropriate weighting mechanism when multi-source mobility datasets are being fused. Therefore, our motivation stems from the need for a novel framework to effectively integrate and learn multi-source data dynamically with domain knowledge for optimal multi-view mobility modeling.

3. Proposed Integrative Methodological Framework

To optimize multi-view mobility modeling, we propose an integrative three-step methodological framework, as shown in Figure 2. The framework combines multi-source mobility data for a more representative picture of travel behaviours, then integrates it with existing domain-knowledge to minimize data requirement while enhancing explainability, and finally conforms the mobility modeling to incremental

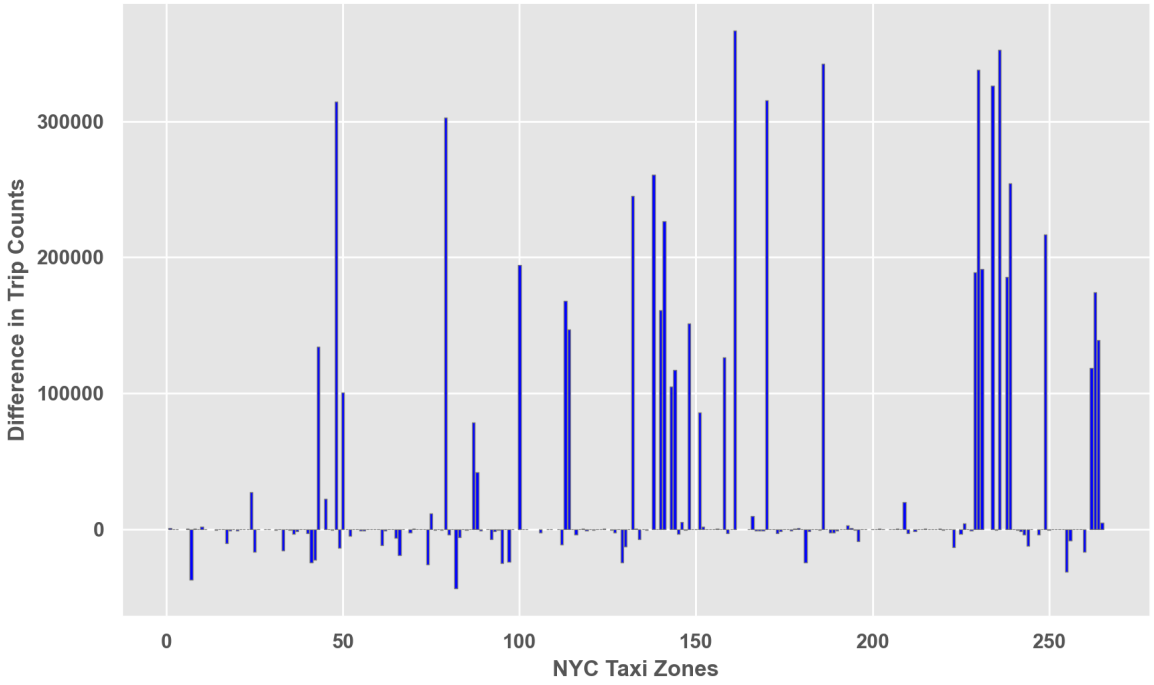

Figure 1: Difference in trip counts by NYC taxi zones

learning, adapting to ever-changing urban mobility conditions. Each component is briefly overviewed in the following subsequent sections.

3.1. Multi-Source Data Fusion

In the context of multi-view learning, data fusion captures the comprehensive view of urban mobility patterns, and also diffuses the limitations that come along with relying solely on single-source data. However, most existing works make general assumption of the multi-source data when it is being fused. For example in [13] , Zhang et al. (2014) used the transport data to simply complement modeling primarily based on cellphone data, and later in [8], Zhang et al. (2017) adjusted and treated the two kinds of data sources equally, as two independent views. This is bound to integrate the data in proportions that don't reflect the true representation of each view. Increasingly, authors in this field [9] [11] [14], acknowledge that data fusion requires a full understanding of each dataset and their efficient utilization, which is currently still limited. In our proposed framework, the characteristic context of multi-source data is leveraged to workout an effective weighting scheme upon which data fusion is done. This will ensure spatial-temporal patterns are realistically and dynamically captured.

3.2. Data – Knowledge Integration

It is widely known that data-driven models are only as good as the quantity and quality of data they are trained on. Given issues relating to incompleteness, inconsistencies, sparsity and uncertainity, it is often difficult to acquire large amounts of quality mobility data. However, integrating empirical data with domain-specific knowledge often validates the data, enhancing model reliability and interpretability, which is still an open challenge in human mobility research. To the best of our knowledge, very few mobility modeling approaches exist in literature today, which incorporate domain-specific knowledge. In MobTCast [15], the influence of semantic, social, and geographical contexts is incomporated with historical data to tackle the sparsity problem, while predicting POIs. Likewise in [16], a framework was proposed to leverage a knowledge graph to accommodate the influence of users, locations and semantics, in next location recommendation. Meanwhile in [17], authors proposed an MaaS framework

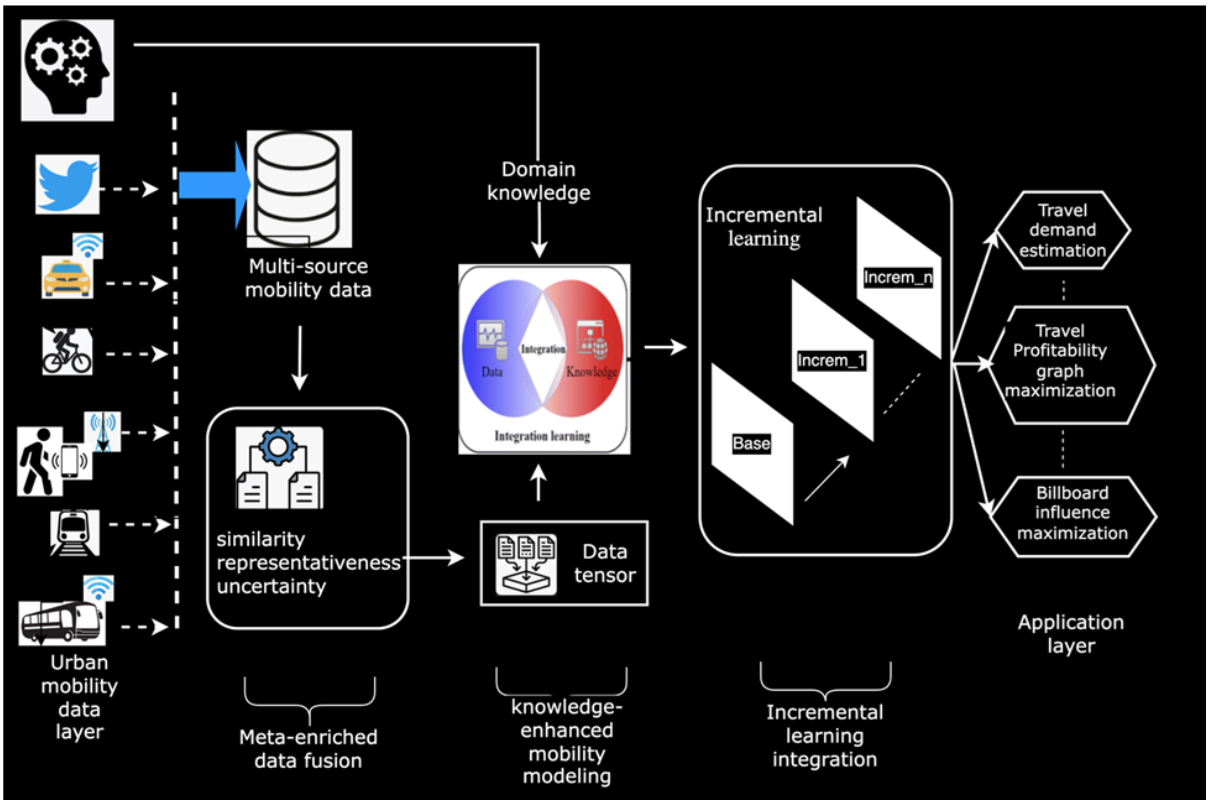

Figure 2: Integrative methodological framework for multi-view mobility modeling

to combine multi-source data with an understanding of travelers' contexts to present them personalized and explainable services based on their preferences. Challenges withstanding, their framework aims to achieve the possibility of providing the right information to the right user with understandable explanations. The common principal about these methods is the integration of empirical data with one or more kinds of contextual understanding of the urban environment. Whereas existing efforts have focused on the influence of somewhat stable-static contexts, our integrative framework goes a step further to incorporate situational contexts which is rather a dynamically changing context.

3.3. Incremental Learning

Urban mobility is highly dynamic, with context and patterns that typically change from time to time. Conventional data-driven models, though they have shown considerable performance, they often struggle to keep at per with evolving urban environments, where mobility patterns keep on changing due to changing policy, transportation systems, road networks, weather conditions, and events. As an emerging approach, Incremental Learning (IL) which enhances adaptability of models by learning continuously from new incoming data streams, is poised to address these challenges suffered by static models. In our framework, the knowledge-enhanced mobility model will be tuned to continuously update on only new incoming data, as and when it becomes available, while retaining previously learned knowledge.

4. Case Study: Urban Mobility in New York City

In the previous section, we give an overview of the proposed integrative three-step methodological framework. **Note** however that in the present work, the focus is to preliminarily use the case of New York City (NYC) to analyze the underlying characteristics of multi-source data, setting up an optimal foundation for effective data fusion and subsquent multi-view mobility modeling in the line of the

proposed framework.

4.1. Study Area

In this work, we characterize and compare mobility flows extracted from NYC's taxi and Citi bike trip records. NYC is the largest city in the United States, known for its fast pace, dynamism and cultural diversity. It continues to be a global hub for finance, culture, commerce, and innovation. It covers a total area of 784 km^2 with an approximate population of 8.5 million people as of 2023. The city is composed of five boroughs: Manhattan, Brooklyn, Queens, Bronx, and Staten Island, each with its own unique character, and it is demarcated into 263 zones, as shown in Figure 3. The city is covered with an extensive and a well-integrated transport system that is comprised of public transit, commuter rails, taxis, and bicycles. A number of researchers have utilized NYC as a case study to investigate different aspects of urban mobility. For example, in [2] F. Miao et al. designed and evaluated the performance of the data-driven vehicle balancing framework on four years' long taxi trip data from NYC, in [18, 19] authors narrowed in on mobility data from NYC in 2019 and 2020 to analyze the impact of COVID-19 on people's mobility, whereas in [20] Dong et al. utilized anonymous mobile phone location and crash report data in NYC to study the association of human mobility and road crashes. We also chose NYC as a case study for this work for three main reasons; (1) open data policy/portal of NYC, (2) shape files for NYC demarcations, and (3) NYC mobility survey report data.

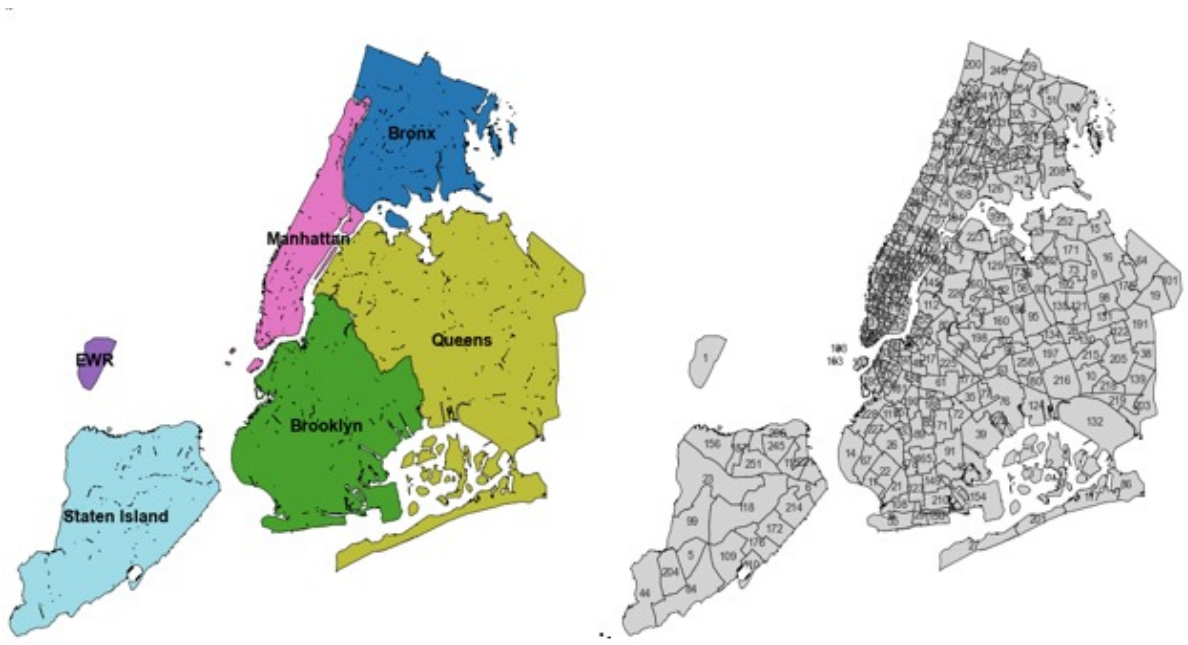

Figure 3: NYC Boroughs & Zones

4.2. Datasets

In this section, the three NYC datasets used in this study, are summarized. The Yellow taxi and Green taxi datasets are provided by the Taxi & Limousine Commission of NYC, while the bike-sharing dataset is provided by Citi bike NYC. For all datasets, each record is a distinctive trip extracted, with an origin, destination, duration, distance and fare among other trip attributes. After pre-processing of the datasets, a total of 10,047,135 trips were successfully extracted from the Yellow taxi, 1,080,844 trips from the Green taxi, and 1,000,000 trips from the NYC Citi bikes. To ensure temporal compatibility, the collection period was the same for all the three datasets, running from April 1, 2017 to April 30, 2017.

4.3. Approach

We characterize multi-source datasets based on three meta-metrics: (1) correlation; and (2) representativeness. We argue that examining and leveraging these three meta-metrics will form a basis for optimal data fusion and subsquently improving the multi-view mobility modelling pipeline.

4.3.1. Trip extraction

Firstly, we extract trips – which is the basic unit of mobility for this study. With all the three datasets having pickup and drop-off attributes, extracting origins and destinations (OD) is simple and straightforward. However, to reduce the size and address errors, irrelevant attributes and errant records were intuitively eliminated right away before any form of feature engineering was done. For example, all records with trip duration < 5 minutes and/or trip distance > 30 miles, were automatically dropped for all datasets.

4.3.2. Correlation analysis

We examine mobility (travel demand) to evaluate whether there exist temporal correlation between Yellow taxi, Green taxi, and Citi bikes trip patterns in NYC. Owing to significant difference in scale among the datasets , we observe in Figure 4 that both Green taxi and Citi bike data temporally lag the Yellow taxi data, but there exist similar mobility trends worth further investigation. To provide fair and meaningful comparison, we normalize the data to capture the hourly variations as a proportion of total mobility flow for each dataset in Figure 5. We then use three different measures (Cosine-similarity, Pearson Correlation, and Spearman Rank Correlation), to explore similarities in their temporal distribution. These measures have been widely used in previous studies to quantify the similarity between two vectors in a multi-dimensional space, depending on the scale and context of the data. In our context, we opted for them because they quantify similarity in patterns regardless the difference in scale.

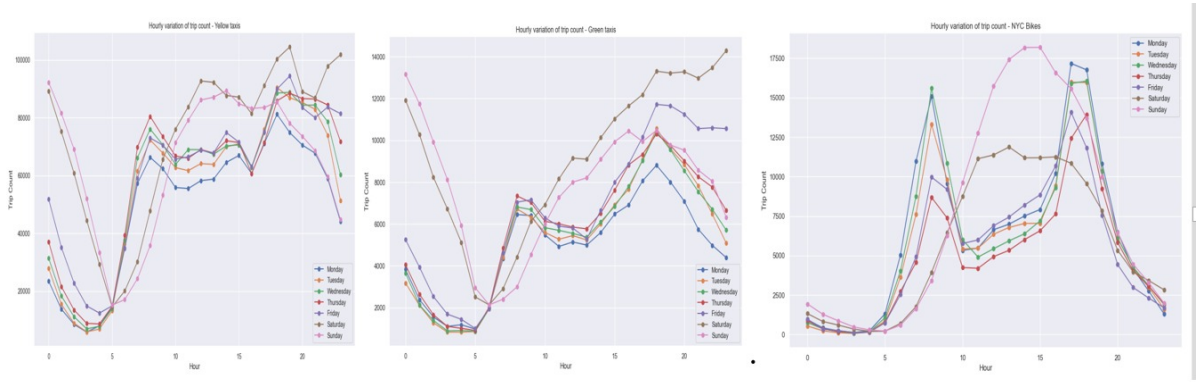

Figure 4: Temporal variation of mobility flow (a) Yellow taxi, (b) Green taxi, and (c) Citi bikes

4.3.3. Representation analysis

Taxi and Bike-sharing mode choices provide NYC residents with certain levels of convenience, timeliness, and flexibility. However, they co-exist and complement other modes in an integrated transport system. It is thus crucial to examine if travel demand extracted from the three datasets reflect their actual respective mobility activity in NYC.

We perform a spatial coverage analysis to ensure that all significant zones are truly represented. For each dataset, we compute travel demand as the proportion of total outgoing trip counts at the zone level. We then use a choropleth map to visualize and compare the spatial distribution of these values, for the three datasets. We then use official statistics from the 2017 citywide mobility survey of NYC, to benchmark patterns derived from our datasets. Particularly we compare the trip proportions extracted

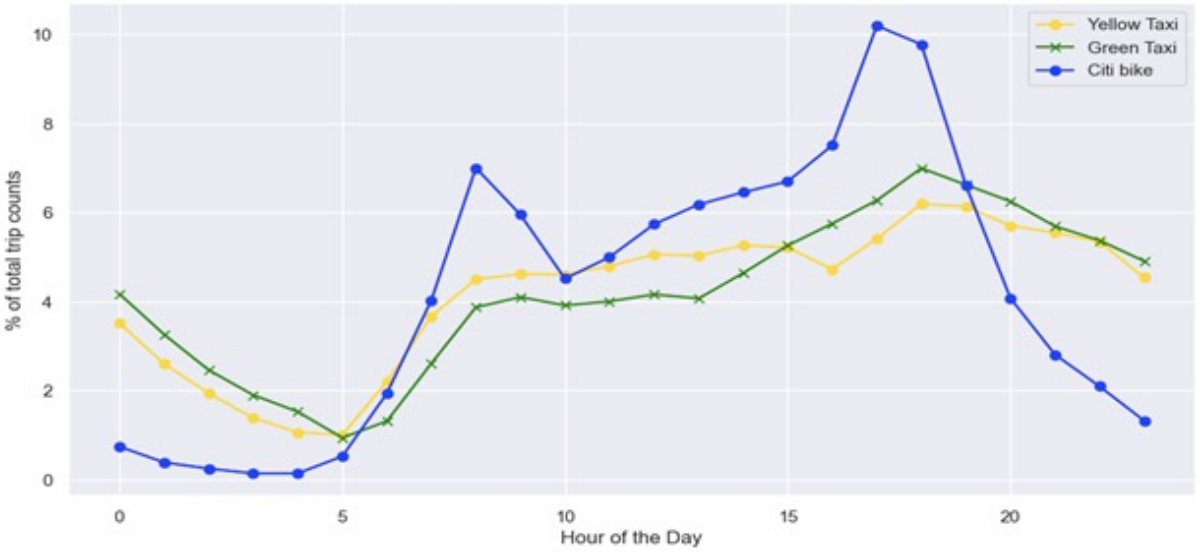

Figure 5: Normalized hourly trip counts for Yellow taxi, Green taxi, and Citi bikes

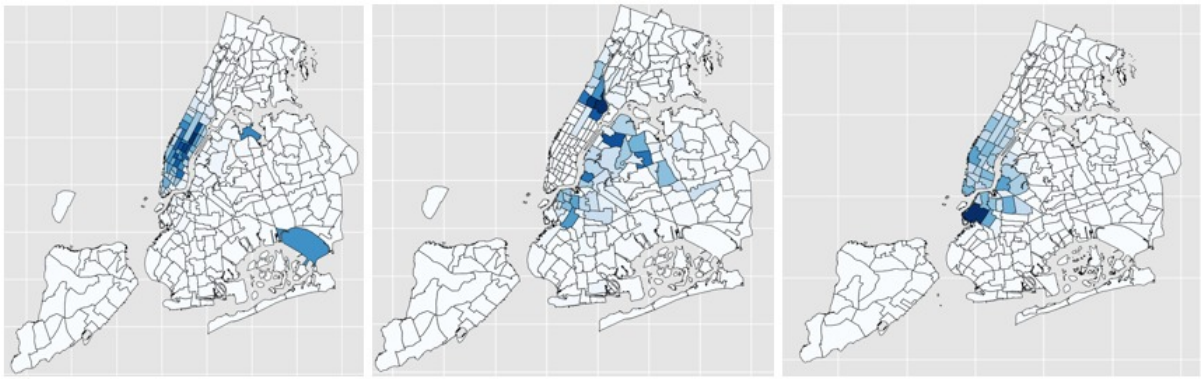

Figure 6: Spatial distribution of trip counts for (a) Yellow taxi, (b) Green taxi, and (c) Citi bike in NYC

from the datasets to the relative trip proportions extracted from the 2017 citywide mobility survey of NYC. And consequently we perform statistical tests on some aspects of the observed trips in comparison with surveyed trips.

5. Results & Discussions

This section summarizes the results obtained along with detailed discussions.

5.1. Correlation

Figure 5 depicts the temporal distribution of normalized travel demand as extracted from Yellow taxi, Green taxi, and Cite bike trip records. Considerable relationships can be observed among the three datasets within the time dimension. For instance, we can see a similar drop in demand from midnight up until 5 a.m., across all the datasets. We also see a shoot-up in demand at 8 a.m. and another one at 6 p.m. To probe these relationships further, Table 1 summarizes results of three well-known measures of similarity against pairs of the three datasets.

From the table 1, we can see that travel demand across all pairs of datasets generally has a positive correlation in the temporal space. Particularly, travel demand patterns from Yellow taxi & Green taxi exhibit a very strong alignment across time, for all the three measures utilized, whereas the correlation

Table 1
similarity measures for hourly distribution of trips extracted from Yellow taxi, Green taxi, and Citi bike datasets

	Cosine-similarity	Pearson Correlation	Spearman Rank Correlation
yellow - green	0.978	0.919	0.906
green - bike	0.899	0.681	0.657
yellow - bike	0.914	0.759	0.710

between green taxi & cite bikes is relatively weak but still considerably above average.

5.2. Representativeness

Figure 6 depicts the spatial distribution of travel demand extracted from the 3 datasets. The hue progressions represent the number of outgoing trip counts aggregated at each zone, divided by the total trip counts for each of the dataset. Yellow taxi & Green taxi are regulated to mainly serve distinct areas of NYC. Yellow taxi primarily serves the core (lower) Manhatten and the major airports, whereas the Green taxi are regulated to operate in upper Manhatten and outer boroughs. We can see how the illustrations in the Figure 6 are compatible with the obvious expectation owing to the above fact. For example in Figure 6.a, we can observe a very sharp concetration of yellow taxi trips in inner Manhatten, JFK Airport, and LaGuardia Airport, in contrast with other zones.

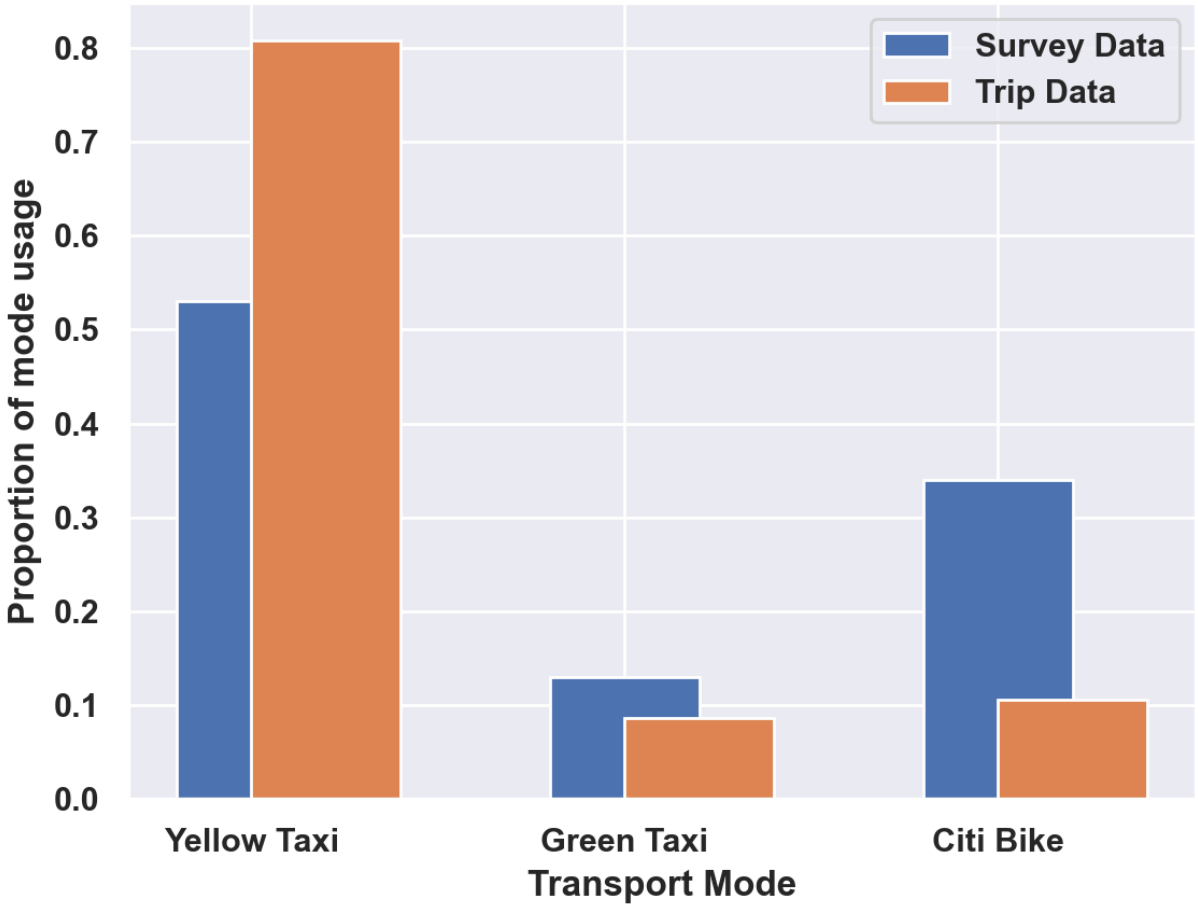

Figure 7: Comparison of Trip data and Survey data

To validate this representativeness, we statistically compare our trip data with the 2017 citywide mobility survey of NYC. Figure 7 illustrates variations in the mode composition as extracted from the trip data and survey data. We observe that among the three modes, Green taxi exhibits a strong fit,

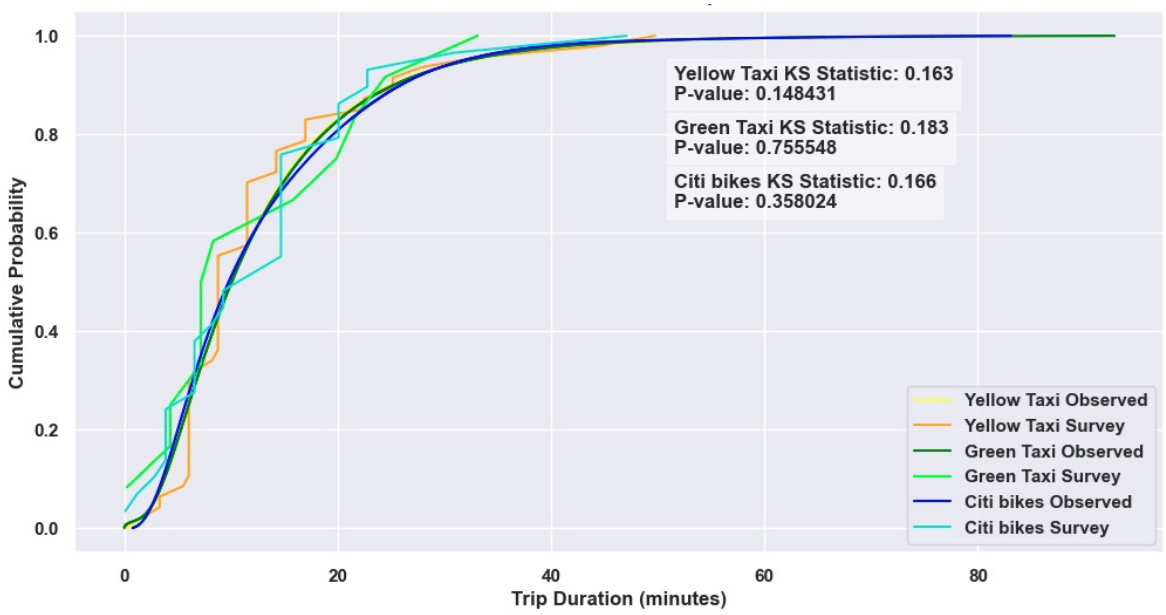

Figure 8: Comparison of cumulative distributions for trip duration

when trip data is compared against survey data. This is further collaborated by the CDF distribution of their respective trip durations as shown in Figure 8. We can observe that though Yellow taxi and Citi bikes both have considerable similar distribution (p-value=>0.05), Green taxi once again exhibit remarkable values of Kolmogorov-Smirnov (KS) test, which demonstrating a strong representativeness of the Green taxi records against the actual mode split as by the 2017 city-wide mobility survey of NYC.

6. Conclusions, Limitations & Future Work

In this study, we proposed a three-step methodological framework for modelling urban mobility from an integration of multi-source data. The preliminary step of our framework involves the systematic characterization of multi-source data as a foundation for a weight-based data fusion. Using NYC as a case study, we characterized three mobility datasets based on two meta-metrics: (1) correlation; and (2) representativeness. Our findings revealed that there exist strong correlation between datasets, which can be leveraged in data fusion. For instance, we see a strong temporal correlation between Yellow taxi & Green taxi, in terms of relative travel demand, which could be used to infer missing data in one dataset using the other. It could also be used in weighted average fusion of the data, where weights are based on the correlation. Also revealed via our findings is the level of representativeness between the respective mobility datasets and the actual urban scenarios. For example we see in our case study how the Green taxi has the most alignment with actual sample data surveyed from the population in the 2017 city wide mobility survey of NYC. This could also be leveraged in the weight scheme to ensure the most representative and reliable dataset contributes most to the data fusion. It should be **noted** that this work does not attempt to examine all the three steps of the proposed framework, but rather to give its overview, and the characterization of the multi-source data to be used as ingredients in the subsequent steps of the framework. One of the limitations of this work however, results from the fact that currently, we only focus on data sources from NYC. If matching multi-source data from other mega cities like Shenzhen and Beijing was readily available, a comparison of mobility trends from different urban cities against the two meta-metrics would further validate the characterization of multi-source data. Another limitation stems from the age of data used (2017), given major mobility shifts that have occurred with the coming Covid-19 pandemic. While very recent mobility data is available at the NYC open data portal, for the interest of measuring representativeness of the data, we wanted to avoid a time difference of data collections between trip records data and the mobility survey data, given that

the city-wide mobility survey data available is of 2017. In our future work, while acknowledging and working around the above limitations, we plan to leverage the findings in this preliminary work, to design an effective weighting scheme for the data fusion. We shall then improvise effective ways to incorporate contextual knowledge in the fused data, and finally employ incremental learning to build an efficient mobility prediction model in line with the proposed methodological framework for different application contexts.

References

[1] Y. Zhou, B. P. L. Lau, C. Yuen, B. Tuncer, E. Wilhelm, Understanding urban human mobility through crowdsensed data, IEEE Communications Magazine 56 (2018) 52–59. doi:10.1109/MCOM.2018.1700569.

[2] F. Miao, S. Han, A. M. Hendawi, M. E. Khalefa, J. A. Stankovic, G. J. Pappas, Data-driven distributionally robust vehicle balancing using dynamic region partitions, in: Proceedings of the 8th International Conference on Cyber-Physical Systems, 2017, pp. 261–271.

[3] H. Rong, X. Zhou, C. Yang, Z. Shafiq, A. Liu, The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency, in: Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 2329–2334.

[4] C. Rodrigues, M. Veloso, A. Alves, C. Bento, Using cdr data to understand post-pandemic mobility patterns, in: EPIA Conference on Artificial Intelligence, Springer, 2023, pp. 438–449.

[5] Y. Li, Z. Ran, L. Tsai, S. Williams, Using call detail records to determine mobility patterns of different socio-demographic groups in the western area of sierra leone during early covid-19 crisis, Environment and Planning B: Urban Analytics and City Science 50 (2023) 1298–1312.

[6] Y. Jiang, X. Huang, Z. Li, Spatiotemporal patterns of human mobility and its association with land use types during covid-19 in new york city, ISPRS International Journal of Geo-Information 10 (2021) 344.

[7] J. C. Wo, E. M. Rogers, M. T. Berg, C. Koylu, Recreating human mobility patterns through the lens of social media: using twitter to model the social ecology of crime, Crime & Delinquency 70 (2024) 1943–1970.

[8] D. Zhang, T. He, F. Zhang, Real-time human mobility modeling with multi-view learning, ACM Transactions on Intelligent Systems and Technology (TIST) 9 (2017) 1–25.

[9] S. Jiang, W. Guan, W. Zhang, X. Chen, L. Yang, Human mobility in space from three modes of public transportation, Physica A: Statistical Mechanics and its Applications 483 (2017) 227–238.

[10] D. Zhang, T. He, F. Zhang, National-scale traffic model calibration in real time with multi-source incomplete data, ACM Transactions on Cyber-Physical Systems 3 (2019) 1–26.

[11] X. Huang, Z. Li, Y. Jiang, X. Ye, C. Deng, J. Zhang, X. Li, The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the us during the covid-19 pandemic, International Journal of Digital Earth 14 (2021) 424–442.

[12] Z. Huang, X. Ling, P. Wang, F. Zhang, Y. Mao, T. Lin, F.-Y. Wang, Modeling real-time human mobility based on mobile phone and transportation data fusion, Transportation research part C: emerging technologies 96 (2018) 251–269.

[13] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, T. He, Exploring human mobility with multi-source data at extremely large metropolitan scales, in: Proceedings of the 20th annual international conference on Mobile computing and networking, 2014, pp. 201–212.

[14] J. Wang, X. Kong, F. Xia, L. Sun, Urban human mobility: Data-driven modeling and prediction, ACM SIGKDD explorations newsletter 21 (2019) 1–19.

[15] H. Xue, F. Salim, Y. Ren, N. Oliver, Mobtcast: Leveraging auxiliary trajectory forecasting for human mobility prediction, Advances in Neural Information Processing Systems 34 (2021) 30380–30391.

[16] Q. Guo, Z. Sun, J. Zhang, Y.-L. Theng, An attentional recurrent neural network for personalized next location recommendation, in: Proceedings of the AAAI Conference on artificial intelligence, volume 34, 2020, pp. 83–90.

[17] E. Rajabi, S. Nowaczyk, S. Pashami, M. Bergquist, G. S. Ebby, S. Wajid, A knowledge-based ai framework for mobility as a service, Sustainability 15 (2023) 2717.

[18] X. Hao, R. Jiang, J. Deng, X. Song, The impact of covid-19 on human mobility: A case study on new york, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4365–4374.

[19] A. A. Rajput, Q. Li, X. Gao, A. Mostafavi, Revealing critical characteristics of mobility patterns in new york city during the onset of covid-19 pandemic, Frontiers in Built Environment 7 (2022) 654409.

[20] N. Dong, J. Zhang, X. Liu, P. Xu, Y. Wu, H. Wu, Association of human mobility with road crashes for pandemic-ready safer mobility: A new york city case study, Accident Analysis & Prevention 165 (2022) 106478.