

Article

Enhancing Travel Demand Forecasting Using CDR Data: A Stay-Based Integration with the Four-Step Model

N. K. Bhagya Jeewanthi ^{1,*} and Amal S. Kumarage ²

¹ Faculty of Built Environment, Tampere University, 33720 Tampere, Finland

² Department of Transport Management and Logistics Engineering, University of Moratuwa, Colombo 10400, Sri Lanka; amal@uom.lk

* Correspondence: bhagya.nagakamkanamge@tuni.fi

Abstract

The growing complexity of urban mobility necessitates more adaptive, data-driven approaches to transport demand forecasting. This study incorporates anonymized Call Detail Record (CDR) data—originally collected for mobile network billing—into the conventional four-step travel demand model to more accurately estimate trip behavior. Employing a stay-based method, significant user locations are identified, and individual mobility patterns are reconstructed. These patterns are then aggregated at the zonal level and validated against a large-scale household survey conducted in Sri Lanka. The proposed framework enables the extraction of origin–destination matrices and supports route assignment using CDR data, demonstrating a strong correlation with traditional survey results. This research highlights the potential of repurposed CDR data as a scalable, cost-efficient alternative to conventional travel surveys for estimating travel demand.

Keywords: CDR data; stay-based approach; activity sequence; load-sharing effect



Academic Editors: Armando Carteni, Andrew Morris and Jo Barnes

Received: 16 June 2025

Revised: 21 July 2025

Accepted: 5 August 2025

Published: 8 August 2025

Citation: Jeewanthi, N.K.B.; Kumarage, A.S. Enhancing Travel Demand Forecasting Using CDR Data: A Stay-Based Integration with the Four-Step Model. *Future Transp.* **2025**, *5*, 106. <https://doi.org/10.3390/futuretransp5030106>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The increasing need for efficiently moving large quantities of goods and people over substantial distances within reasonable timeframes has highlighted the importance of reliable traffic and transportation planning. As a result, there has been a significant focus on planning and implementing transportation initiatives in recent history. Today, transportation investment and effective planning have become an integral part of economic and development strategies. These plans often include developing transit systems, pedestrian and cyclist facilities, and measures to manage transportation demand.

Transportation planning, as the initial step in the implementation of transportation systems, constitutes a significant technical process that heavily relies on computer models and advanced tools to simulate the intricate interactions inherent in transportation system performance. Effective transport planning is vital for enhancing the livability and efficiency of cities, thereby ensuring their preparedness for the challenges of future eras. This intricate transportation planning process often necessitates engagement with a broad spectrum of stakeholders, primarily focusing on infrastructure users. The fundamental action in planning transport initiatives is the measurement and analysis of people's movements. Comprehensively scrutinizing these movements is critical to developing transportation infrastructure and services that operate efficiently while minimizing the burden on end-users.

Typically, information corresponding to human travel activity or movements may be acquired using different approaches, such as survey-based, passive, activity-based, and device-based. Survey-based approaches follow a method of gathering data from people engaged in a survey. Survey methodology targets instruments or procedures that ask one or more questions that may or may not be answered. Researchers conduct statistical surveys to make inferences about the population being studied. Ordinary travel surveys assemble human activity data, preferences, and behavioral data through questionnaires. The veracity of survey-based approaches is widely influenced by the memory of each individual participating in the survey. Additionally, the cost associated with conducting a survey is expensive and laborious. Due to this, the frequency of surveys is low, often leading to small sample sizes and an increased risk of sampling bias. A lower response rate due to response burden is another problem encountered, but certain complex data types, like preferences or opinions, can be gathered easily. Moreover, the existing studies on CDR generate only the screen line flows without the OD (origin–destination) details.

As discussed above, the characteristics of each approach, including data availability, accuracy, scalability, cost, privacy concerns, and information content, vary drastically. Therefore, careful consideration should be given when selecting the data type and acquisition method methodology. In recent decades, researchers have increasingly turned to using time-stamped location data derived from mobile network providers, combining aspects of both passive and active approaches in modeling human behavior. This data type is commonly referred to as Mobile Network Big Data. Technological advancements, including improved data retrieval, processing, and storage facilities, have enhanced the usability of such big data, motivating researchers to incorporate these datasets into their studies.

This study presents a novel approach to integrating CDR data into the four-step travel demand modeling framework (Figure 1). Modeling travel demand typically begins with household surveys to collect transportation data, such as trip frequency, origins, destinations, and times. The initial phase of the traditional transportation model estimates the total number of trips originating from and destined for each zone within a region, classifying them based on their purpose—such as Home-Based Work, Home-Based Other, or Non-Home-Based trips. In the second stage, these trips are processed into an origin–destination (OD) trip matrix, disaggregated by travel modes, such as private vehicles, public transport, walking, and cycling. The final step is route assignment, where trips between each OD pair by mode are estimated and loaded into the transport network to determine the total number of trips for each route [1–3].

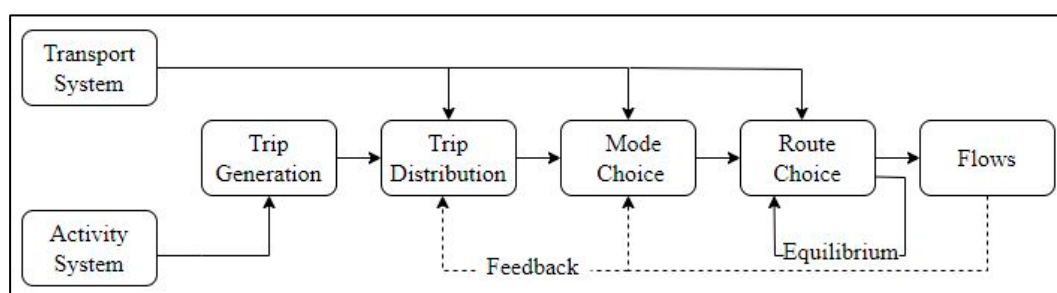


Figure 1. Traditional four-step modeling process.

Within this study, we apply a stay-based methodology to extract significant user locations (e.g., Home, Work, and other) and define trips based on regularity and frequency of cell tower appearances. These trip patterns are then aggregated into origin–destination matrices and validated against a comprehensive household travel survey conducted in Sri Lanka. The final stage involves assigning inferred trips to a road network using a

route choice model, which utilizes the sequence of cell tower connections made during commuting—referred to as en-route cell data—to approximate users' likely travel paths.

By combining mobile network data with traditional modeling principles, this research demonstrates that CDR-based methods can yield comparable accuracy to survey-based models while offering significant advantages in scalability, cost, and temporal coverage.

2. Background and Related Work

Call Detail Records (CDRs) have been extensively used in mobility research to understand individual movement patterns and broader human mobility trends [4–12]. These records—collected passively from mobile phone activity—enable researchers to reconstruct individual trajectories across space and time. Such trajectories, shaped by daily routines and activity participation, define a person's activity space [7,13]. This concept has been central to many mobility studies, where CDR data is used to infer trips, identify frequently visited locations, and analyze travel behaviors. The literature using CDR data to model human mobility broadly falls into three categories: OD matrix estimation, significant location identification, and route assignment.

2.1. Origin–Destination Matrices

A large body of research has explored how to derive OD matrices from raw CDR traces. Fekih et al. [14] proposed identifying stationary periods in user traces to determine activity zones, forming the basis for trip detection. Similarly, Mamei et al. [15] inferred OD pairs from sequences of CDR appearances, relying on temporal and spatial transitions. These approaches generally use a trip-based model—segmenting travel behavior based on location transitions within specific time windows.

Other studies, such as Bwambale et al. [16], introduced latent demographic modeling, weighting trip generation by inferred socio-demographic attributes derived from mobile phone usage characteristics. While such methods enhance behavioral realism, they often lack direct validation. Meanwhile, studies in Boston, San Francisco, and Dhaka [5,17] constructed tower-level transient OD matrices, converting them to node-to-node networks for routing, yet often without addressing accuracy at the individual level.

2.2. Mobility Pattern Recognition and Significant Location Identification

This category of research focuses on identifying anchor locations (e.g., Home, Work) around which users organize their travel. These locations are typically inferred by analyzing appearance regularity at specific cell towers. Luo et al. [18], for example, combined CDR and road network data to reduce tower-level noise and extract Home/Work locations using trajectory regularity. Other studies used clustering algorithms to group spatially adjacent towers [19] or time-weighted frequency scores to detect commonly visited places [20,21].

Leng et al. [22] proposed a geo-temporal matrix approach, using eigen decomposition to identify recurring spatiotemporal patterns. These methods often assume that trip behavior is governed by routines, but they may apply uniform thresholds across all users, potentially missing individual variation in location behavior.

2.3. Route Estimation from CDR Data

CDR-based route estimation involves mapping inferred OD pairs onto a transport network. The first step typically involves generating a set of feasible routes, often using OpenStreetMap data [23,24]. Then, a flow distribution algorithm assigns trips to these routes. One common method is the use of defined cell paths, where sequences of cell tower connections are used to approximate the chosen path [25].

2.4. Research Gap and Objectives

Despite the breadth of CDR-based mobility studies, key limitations persist:

- OD estimation and route assignment methods often ignore individual-level behavioral variations, relying instead on aggregated transitions.
- Route choice is typically modeled based on tower frequency or flow density, without considering parameters such as appearance regularity, call frequency, or user-specific trip patterns.
- Few studies validate CDR-derived mobility outputs against traditional travel survey data at a granular (zonal or individual) level.
- This study addresses these gaps by introducing three key innovations:
- A user-specific regularity model that detects significant locations (e.g., Home, Work, and other) using temporal and frequency-based indicators.
- An individual caller-based OD estimation approach, which constructs personalized travel demand profiles rather than relying on aggregated cell transitions.
- A route assignment methodology using en-route cell sequences, which calculates route alignment probabilities for each user based on observed call paths—thereby modeling route choice with individual-level granularity.

Through this framework, we demonstrate that privacy-preserving CDR data can be repurposed as a scalable and behaviorally nuanced alternative to traditional survey-based methods for travel demand estimation.

3. Study Area and Data

This study uses Call Detail Records (CDRs) from a mobile operator in Sri Lanka, provided by LIRNEasia, a regional ICT policy think-tank. The data is pseudonymized to ensure user privacy. Each record includes a user ID, call time, cell tower connection, and call duration, enabling geographic location association. The dataset covers 400,000 callers over the month of June 2013, totaling 89 million call records, focusing on the Western Province of Sri Lanka.

The primary data source for validation is transportation-related data collected through the Household Visit Survey (HVS) conducted in 2013 as part of the CoMTrans study in the Western Province of Sri Lanka [26]. The Household Visit Survey (HVS) data includes transportation information, such as socio-demographic records, travel times, trip purposes, and travel modes. This data is utilized to validate the results from the CDR trip analysis. The survey used a face-to-face interview method and covered 44,000 households, with a response rate of 4%. The HVS includes detailed socio-demographic data, travel times, trip purposes, and travel modes.

To ensure comparability between the CDR-based outputs and the HVS data, CDR tower locations were mapped to corresponding administrative zones (districts and Divisional Secretariate Divisions: DSDs). Aggregation was performed to align the granularity of CDR data with the travel zones used in the household survey. Differences in temporal coverage and behavioral representation were addressed by validating trip distributions and route assignments at both the macro (district) and micro (DSD) levels. Statistical methods, including correlation analysis and outlier tests, were applied to evaluate the consistency between the two datasets.

4. Methodology

Figure 2 outlines the methodology, starting with a 400,000-sample CDR dataset collected over a month. Geospatial data on cell tower locations and administrative boundaries are foundational inputs. The Household Visit Survey (HVS) dataset is integrated for validation. Initially, the user coordinates with timestamps to reveal digital movements. Records

are categorized into Home/Work and other points based on location patterns. Trips and stays are delineated from the dataset. Results are validated against traditional data at administrative zonal levels, showcasing the refined CDR-based Travel Demand Estimation Model in four stages.

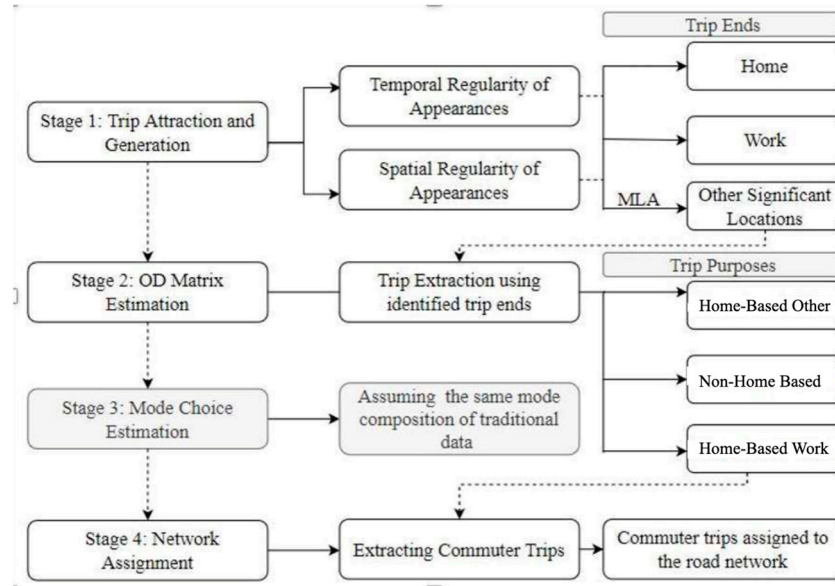


Figure 2. Four-step modeling process using CDR data.

Before demand modeling, filtering noise from tower-to-tower call balancing is crucial. This balancing creates false movements, known as the load-sharing effect. To correct for the load-sharing effect—a phenomenon where users appear to move due to operator-based call balancing rather than actual physical movement—this study applies a speed-based filtering technique. The method assumes that speeds exceeding 40 km/h between two consecutive cell tower connections within a short time interval are implausible under typical urban conditions and are likely artifacts of load balancing. By calculating speeds based on tower centroid coordinates and elapsed time, the approach identifies and removes these anomalies. A range of speed thresholds (10–90 km/h) was evaluated using Divisional Secretariate Divisions (DSDs) and Kullback–Leibler divergence to compare CDR-inferred locations with household travel survey data. The optimal threshold of 40 km/h minimized distributional error, supporting its use as a cutoff for filtering out spurious records. This ensures that only legitimate movement patterns are retained for downstream mobility analysis and model estimation [27,28].

4.1. Stage 1: Trip Attraction and Generation

With CDR data, routes are identified by pinpointing anchor locations first. Origins and destinations are determined based on consistent appearances at these points. Frequent locations, like Home and Work, alongside other significant spots, serve as trip generation and attraction points. Movement between these points signifies a trip. Identifying these locations is crucial as they generate more calls. Individuals’ behavior divides cell towers into regular and irregular ones, with regularity varying individually. Once identified, regularly visited cells, such as Home or Non-Home, are labeled as significant locations.

Identifying Home and Work Locations

The location most visited in the morning is referred to as “Work”, while the most frequent nighttime location is identified as “Home” [5,6]. To more accurately determine the Home location, we compare the top nighttime locations during weekdays and weekends.

If these locations matched closely, they are reliably labeled as Home. Specifically, the cell tower most often connected to between 8:00 PM and 4:00 AM is designated as the Home location. In contrast, Work locations are defined as places users regularly visit during weekday daytime hours but rarely during weekends. These locations are assumed to represent workplaces where users spent long periods. Therefore, the Work location is determined by identifying the cell tower most frequently accessed during typical office hours (10:00 AM–4:00 PM) on weekdays.

It is important to note that the “Work” classification includes both workplaces and educational institutions, such as schools and universities. Due to the anonymized nature of CDR data, it is not possible to distinguish students from working individuals. As a result, frequent daytime locations during office hours (10:00 am–4:00 pm) are collectively treated as Work/Study locations.

Once the Home and Work locations are identified, it is critical to specify significant locations that are neither Home nor Work. This study uses the appearance regularity of the cell locations to explore the above.

Identifying Significant Other Locations: Home–Work

This study assesses how consistently users visit each cell tower location using three factors: (a) consistency based on the time of day, (b) consistency based on the day of the week, and (c) consistency based on the day of the month. A day is divided into 10 time intervals (T_1 to T_{10}), such as $T_1 = 6–8$ AM, $T_2 = 8–10$ AM, and so on up to $T_{10} = 4–6$ AM. The regularity of a user (n) being present at a specific cell (x) during time slot T_i is proportional to the following:

1. Total number of days user n visited cell x during T_i .
2. Number of days per week the user visited cell x during T_i .
3. Number of days per month the user visited cell x during T_i .

Hence, the user’s regularity at cell x during T_i is proportional to $a \times b \times c$.

After calculating regularity for all cells and time intervals, the cell with the highest regularity is identified as a significant location. This process is personalized for each user. To accomplish this, this study uses Mathematical Linkage Analysis (MLA), a graph theory-based technique that evaluates how closely a user’s movement pattern aligns with an ideal regular pattern across cells [29]. It calculates the R-squared value between actual and ideal patterns for different cell groupings. The configuration with the highest R-squared identifies the most significant locations—excluding Home and Work. These significant places form the first stage of a four-part model, where Home, Work, and other major locations act as points of origin and destination for user movement, defining their travel patterns.

4.2. Stage 2: OD Matrix Estimation

The goal of the second stage is to match the trip origins and destinations identified in the previous step to form origin–destination (OD) pairs. This results in a trip matrix (T_{ij}) categorized by trip purpose. The classification of trip purpose depends on the nature of the origin and destination locations. If a trip starts at Home and ends at Work (or vice versa), it is labeled a Home-Based Work trip. If the trip starts or ends at Home but the other location is a significant place other than Work, it is classified as a Home-Based Other trip. Trips where neither end is Home are categorized as Non-Home-Based trips.

Once the trip purposes are defined, they need to be aggregated into Traffic Analysis Zones, which comprise the study area. Traffic Analysis Zones can be selected to encompass the expected resolution level. Considering the coverage area and complexity of overlaying, this study matches the cell towers to the district and DSD levels, such that the OD matrices are developed at the intra-district and inter-DSD levels.

4.3. Stage 4: Network Assignment

This section analyzes the route choice behavior of commuter trips, particularly Home-Based Work trips, inferred from the stage 2 OD flows. Trips are further analyzed using defined statistics to generate route choice probabilities. “En-route cells” and “cell paths” are introduced to assess the nature of route choice. En-route cells represent cell connections made during commuting, extracted from the last call at Work and first call from Home, and vice versa. A distance filter is applied to qualify cells as en-route cells. The network is interpreted using cell paths overlaid with Voronoi shape files to demonstrate sequences of en-route cells as routes. Table 1 presents the definitions of trip statistics and the method used to derive route choice based on those statistics.

Table 1. Defining trip statistics.

Disaggregate Measures—Per Trip		Aggregate Measure—Per Individual
(a) Alignment of a trip to the defined route—(r)	(b) Certainty in route alignment prediction of a trip	(c) Contribution of a user to route choice for the route (r)
The number of en-route cells that fell along the defined cell path (A) was divided by the total number of cells along the cell path (B). This generated the probability of alignment to the defined route for each user’s trip (M). $M = P(A)/(B)$	It was measured using the no. of connections made to the cells along the route in each trip. The no. of connections was used as a weighting factor such that higher weightage is assigned when the no. of connections is high (WC).	Two measures were used to quantify the contribution made by each user to the route inference. 1. The output probability of measure (a) and (b), which is the probability of a user selecting a defined route for commuting. 2. Weightage was derived by dividing the trips with evidence (T) from the total trips (Y).
The probability of user X selecting the defined route for commuting, $n =$ no. of trips $P(X) = \frac{\sum_{i=1}^n M_i \times WC_i}{\sum_{i=1}^n WC_i}$		Contribution of ser X for route (r) $P(X_{Final_R(r)}) = P(X) \times (W/Y)$
The final route choice model was derived by multiplying the probability along a specific route by the actual number of days the caller has traveled to Work. The exact process was carried out for each user; then, the proportion via the considered route was calculated as in the equation below. It was assumed that the same route selection probability exists for trips in which the en-route cells are not available. Figure 3 presents the graphical representation of the route choice.		
$(m) =$ No. of users among a particular DSD pair. $(r) =$ No. of routes among a particular DSD pair. $P(X_{Final_R(r)}) =$ Probability of user X selecting route (r) for commuting. $(T_i) =$ No. of days the user had traveled to Work. The final route selection probability can be shown in the equation below, $\text{Percentage of Trips made via route } n = \frac{\sum_{i=1}^m P(X_{Final_R(n)}) * T_i}{\sum_{r=1}^r [\sum_{i=1}^m P(X_{Final_R(n)}) * T_i]}$		

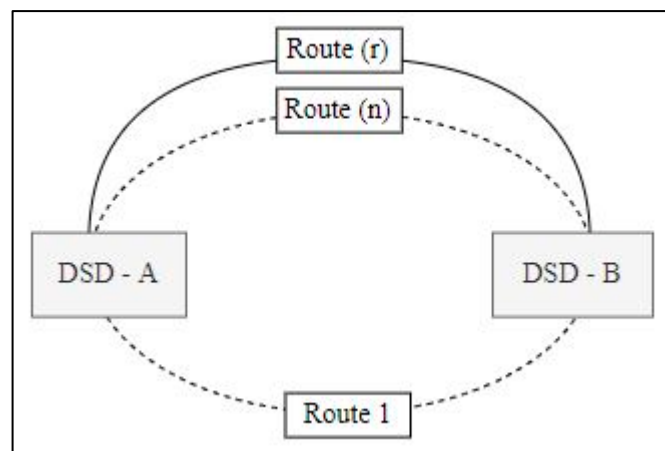


Figure 3. Route choice representation.

The main output of this procedure is the percentage of commuting trips assigned to the road segments within the study area. The commuting trips thus generated are then

expanded to actual values and assigned using STRADA software to the road network by following the standard process in the Comtrans network assignment. The User-Equilibrium method is employed in assigning private vehicle trips, and the transit assignment method is used for public transit trips. It is crucial to note that the CDR trip assignment assumes that the mode composition observed in the model outputs prevails.

Route assignment with CDR initially uses trips with en-route cells and expands to calculate the total work trips, assuming that trips without en-route also behave like trips with en-route.

Accordingly, for all DSD pairs, O_iD_j , where $i \neq j$,

$$\text{With CDR : } \mathbf{1} = \sum_{n=1}^m \left[\frac{R_{n_CDR}}{[\sum_{i=1}^n R_{n_CDR}]} \right]$$

$$\text{With HVS : } \mathbf{1} = \sum_{n=1}^m \left[\frac{R_{n_Strada}}{[\sum_{i=1}^n R_{n_Strada}]} \right]$$

R_{n_CDR}/R_{n_STRADA} —No. of trips via route n with CDR/Modeled.

m —No. of routes identified for the considered DSD pair ($n = 1, 2, 3, \dots, m$).

i —Origin DSD (1, 2, 3, ..., 47).

j —Origin DSD (1, 2, 3, ..., 47).

4.4. Summary of Methodology

Figure 4 below summarizes the overall methodology, distinguishing between the traditional and CDR methods.

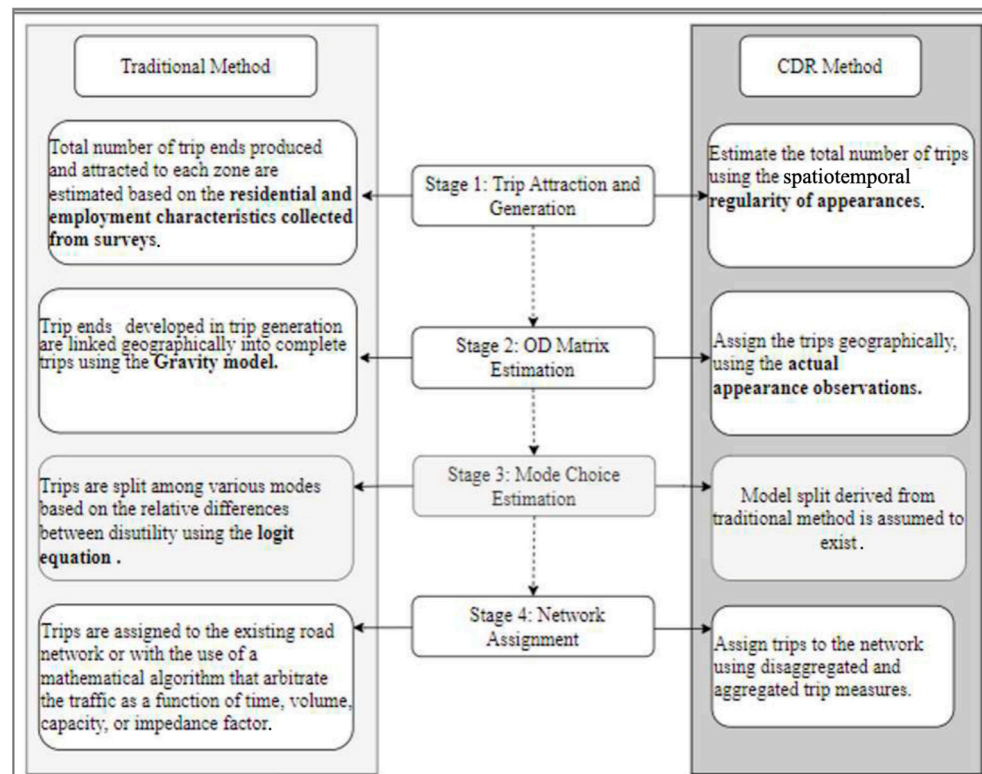


Figure 4. Comparison between the CDR-based and Traditional Travel Demand Estimation Models.

CDR data offers significant insights into mobile phone users’ spatial and temporal patterns, aiding the estimation of trip origins and destinations. Table 2 and Figure 4 highlight the complementary strengths of each approach and motivate the integration of CDRs into planning frameworks where traditional surveys may be logistically or financially

constrained. These records, capturing call and text message details, facilitate the analysis of communication patterns, such as call/text frequency, duration, and timestamps. Such analysis contributes to inferring trip purposes and rates, enhancing the accuracy of trip generation models by estimating trips from various zones, and identifying travel behavior variations by time or day.

Table 2. Comparison between traditional and CDR-based models.

Feature	Traditional Model	CDR-Based Model
Data Source	Household travel surveys	Mobile network (CDR) data
Temporal Coverage	1–2 days (snapshot)	Continuous (e.g., 1 month in this study)
Spatial Resolution	Traffic zones or administrative units	Cell tower level (finer spatial granularity)
Frequency of Updates	Infrequent (every 5–10 years)	Frequent or real-time
Cost and Logistics	High (manual effort and cost)	Lower (automated, passive data)
Behavioral Detail	Self-reported, subject to bias	Inferred from actual usage patterns
Privacy Sensitivity	Voluntary disclosure	Requires strong anonymization safeguards

In the trip distribution phase, CDR data elucidates mobile phone users' destinations, enriching the understanding of travel patterns and modeling trip flows between zones. This data aids in assessing the probability of trips between different origins and destinations, leveraging the spatial distribution of communication events to pinpoint prevalent travel patterns and flow dynamics. This approach helps estimate trip attractions and generations for specific areas, highlighting temporal variations like peak periods and congestion zones.

5. Experimental Details and Results

This section outlines the outputs generated by our CDR-based Travel Demand Estimation Model. It begins with the development of trip tables and a comparison of origin–destination (OD) matrices with those derived from conventional travel surveys. We then present road network assignment results and validate them against traditional traffic assignment approaches.

5.1. Trip Tables and Survey Comparison

To generate OD matrices, a processed sample of Call Detail Record (CDR) data was analyzed. The analysis included identifying user locations categorized as Home, Work, and other significant places. Using CDRs, OD pairs were determined between cell towers; however, to align with traditional datasets, these tower-level locations were mapped to administrative units. In the study area, 763 mobile towers were mapped to 48 Divisional Secretariat Divisions (DSDs) and 3 districts, which were defined as OD-generation nodes.

Trips were identified by changes in consecutive cell tower locations and assigned to OD matrices at both the district and DSD levels. When a cell tower was located near administrative boundaries, trip generation and attraction counts were proportionally distributed based on the overlapping area between zones.

The graphs (Figures 5 and 6) compare CDR-derived trip types—Home-Based Work (HBW), Home-Based Other (HBO), and Non-Home-Based (NHB)—with traditional Household Visitor Survey (HVS) data. All trip types show a strong correlation with the HVS results at broader geographic levels, though accuracy declines slightly at more granular (DSD) levels. HBW trips, due to their more routine nature, demonstrated the highest agreement.

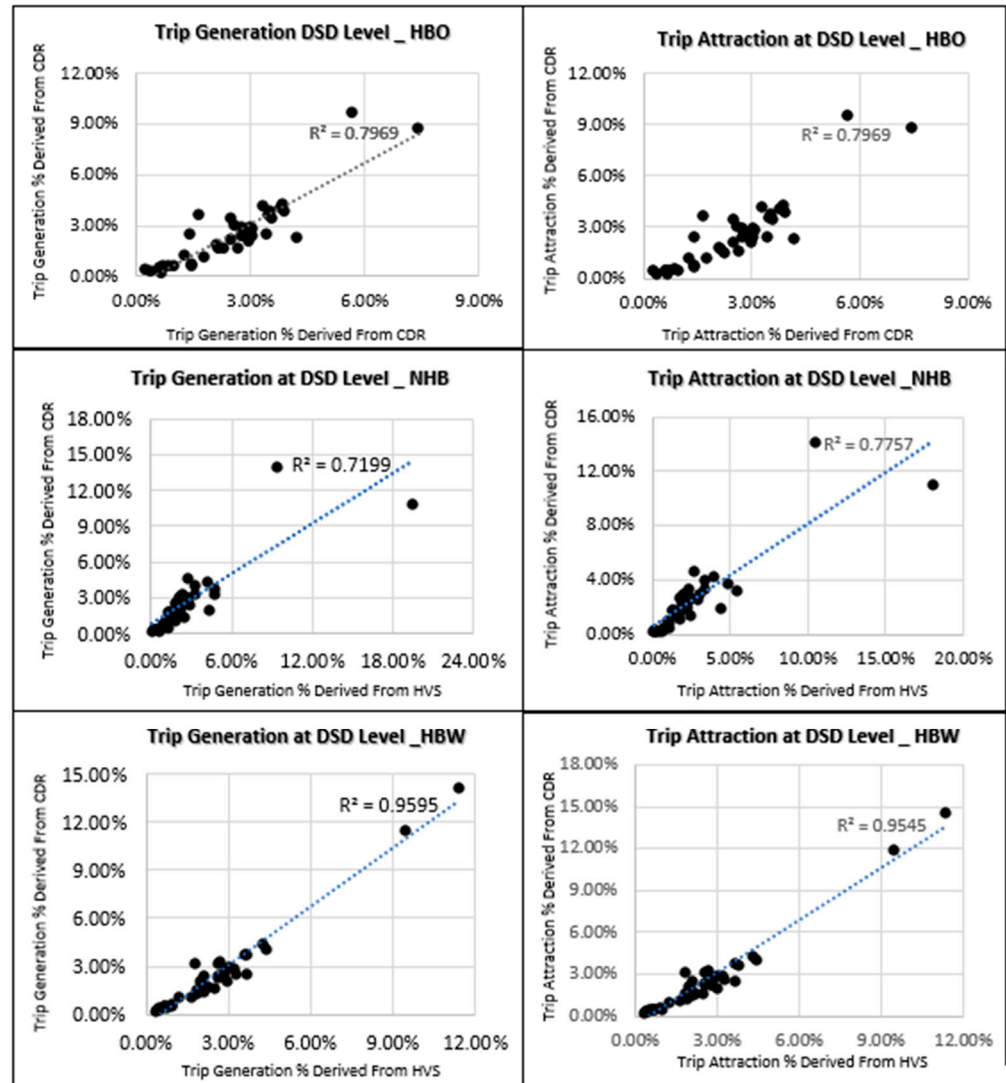


Figure 5. Trip distribution comparison at the DSD level.

At the district level, we observed the following:

- HBW trips had an R^2 of 97% and a standard error of 19%.

At the DSD level, we observed the following:

- Trip generation had an R^2 of 80% (standard error: 38%).
- Trip attraction had an R^2 of 85% (standard error: 33%).

These decreases in accuracy at more disaggregated levels are mainly due to the difficulty of accurately aligning cell tower coverage areas with smaller administrative travel zones.

An outlier analysis using R identified Colombo and Thimbirigasyaya as anomalies in the HBO and NHB categories, likely due to proximity-related boundary errors. When these two DSDs were combined into a single unit, the correlation improved notably.

High-density urban zones like Colombo tend to have overlapping tower signals and frequent handoffs, contributing to errors in trip allocation. While these effects are diluted at the district level, they introduce significant variance at the DSD level. Adjusting for such anomalies is key to improving model precision in future applications.

To supplement Figures 5 and 6, Table 3 presents the actual and expanded HBW trip counts at the district level. The CDR data was scaled to reflect the actual population, as per the model’s methodology. Despite differences in absolute values, the CDR-based

OD patterns align well with the HVS data, particularly for high-volume OD pairs like Colombo–Colombo and Gampaha–Gampaha.

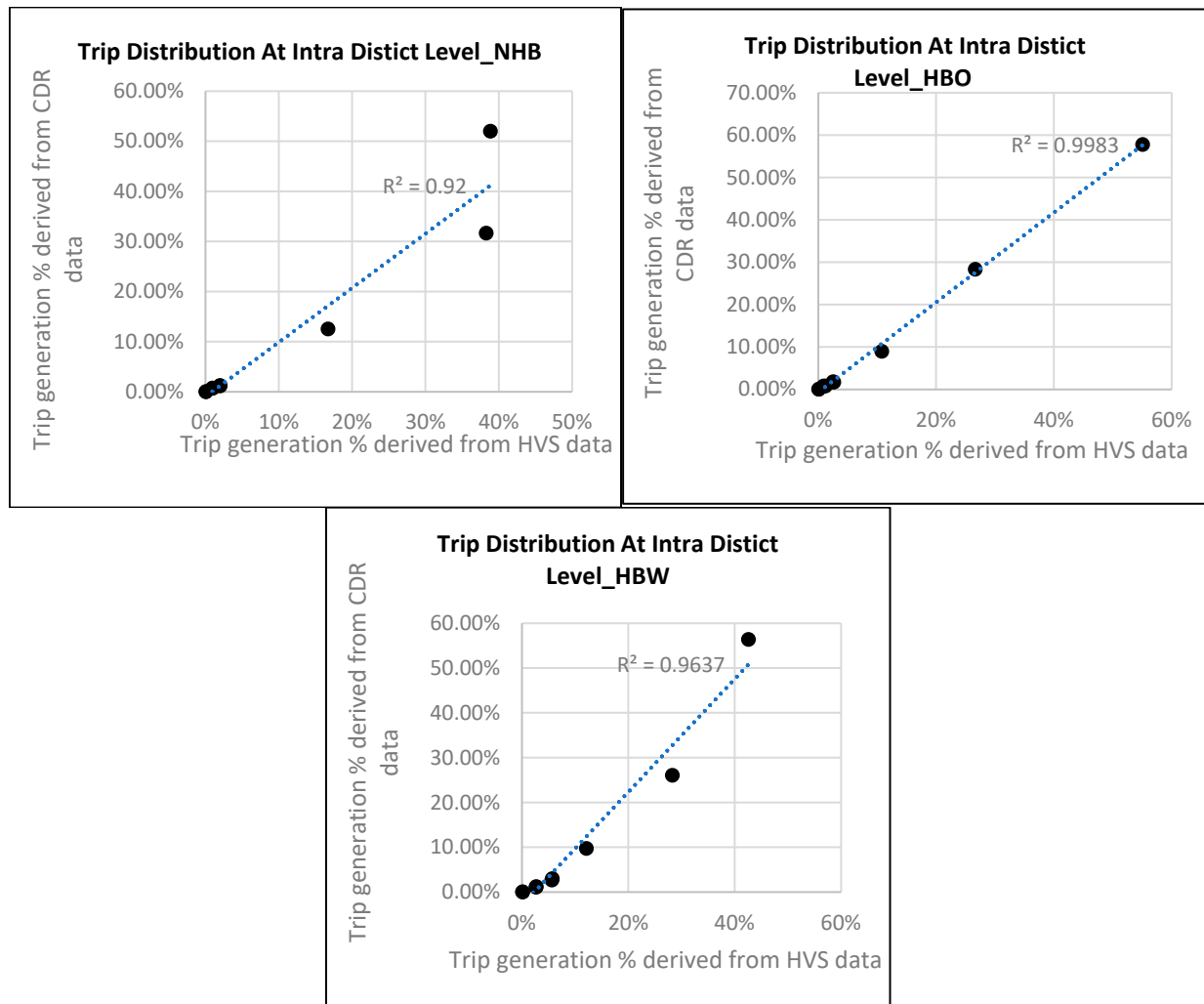


Figure 6. Trip distribution comparison at the district level.

Table 3. Expanded data.

OD Pair (District Level)	CDR Data (Expanded)	HVS Data
Colombo–Colombo	290,039	1,150,559
Colombo–Gampaha	18,658	152,107
Colombo–Kalutara	10,725	71,792
Gampaha–Colombo	21,538	154,907
Gampaha–Gampaha	148,854	763,714
Gampaha–Kalutara	206	3388
Kalutara–Colombo	12,076	72,925
Kalutara–Gampaha	73	3466
Kalutara–Kalutara	68,374	327,588

5.2. Road Network Analysis

Figure 7 illustrates the validation of route choices produced by the CDR-based model against traditional assignment methods. Figures 8 and 9 display line charts highlighting movement trends between the DSDs.

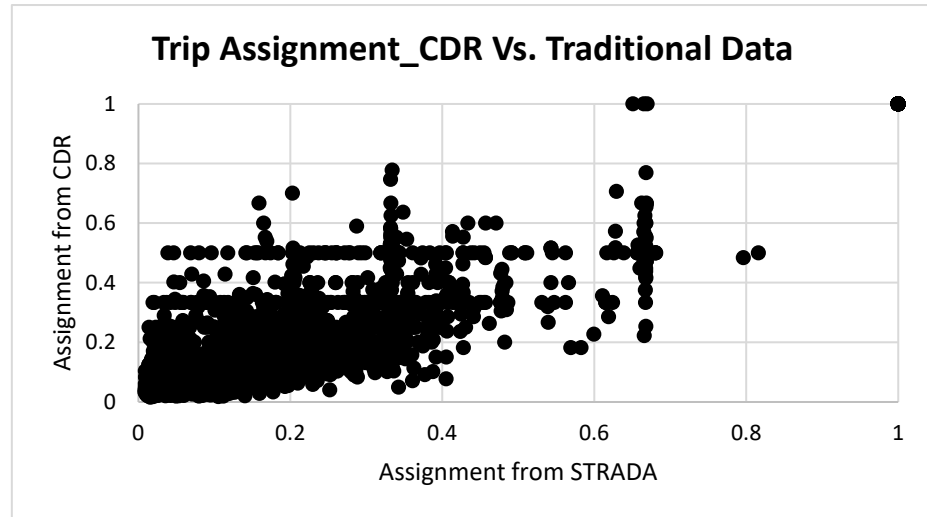


Figure 7. Trip assignment: CDR vs. traditional data.

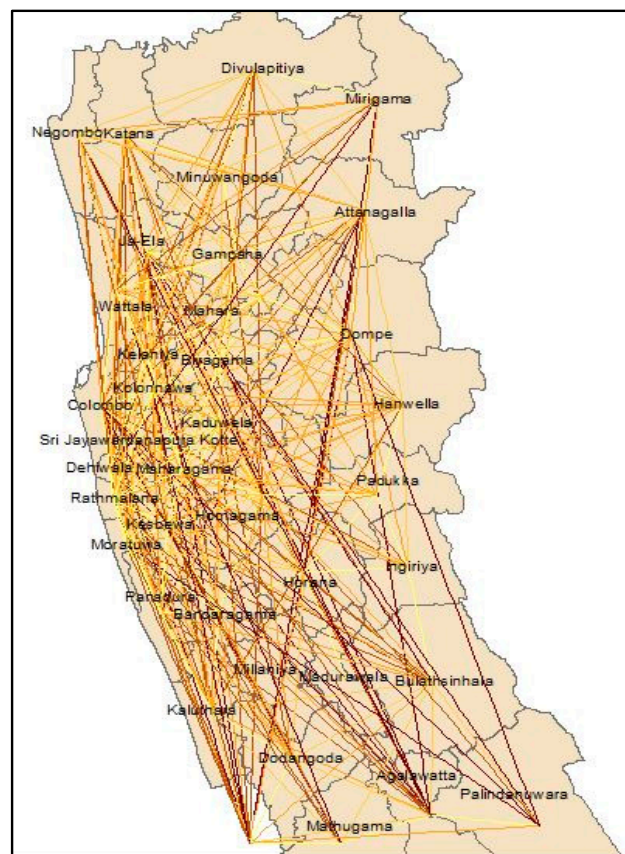


Figure 8. Route assignment from CDR data.

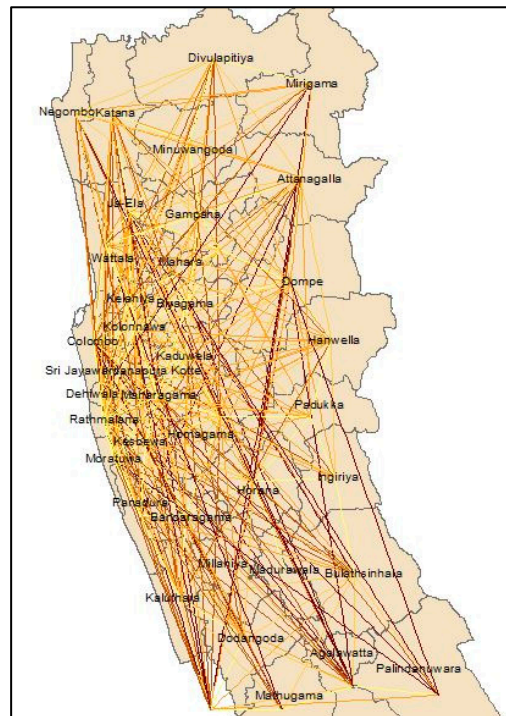


Figure 9. Route assignment from traditional data.

The correlation between the STRADA and CDR was 0.83, which was statistically acceptable. However, some outliers could be detected by visual inspection. When the data was tested using an outlier test, a few DSD pairs, namely, Bandaragama–Kelaniya, Dodangoda–Kaduwela, Gampaha–Dehiwala, Homagama–Panadura, Horana–Kelaniya, Kolonnawa–Dompe, Mirigama–Dehiwala, Negombo–Ratmalana, Padukka–Attanagalla, and Panadura–Kolonnawa, were detected at the initial iteration. These DSD pairs had a significantly lower number of Work trips compared to the other pairs.

A hypothesis test was carried out to determine the significance of the correlation and, based on that, whether the null hypothesis could be rejected in favor of the alternative. The p -value at 5% significance was 0.00001. Since the p -value is smaller than the significance level ($\alpha = 0.05$), we can reject the null hypothesis in favor of the alternative and conclude that the correlation is statistically significant or that there is a linear relationship between the two variables in the population at the α level.

6. Conclusions

This study presents a comprehensive travel demand modeling framework that leverages mobile phone Call Detail Record (CDR) data to enhance and potentially replace traditional survey-based approaches. Compared to conventional household travel surveys, CDR data provides broader temporal coverage—capturing multiple days of user movement, including weekends—and offers scalable, cost-effective, and frequently updated mobility insights.

A novel methodology is introduced to construct origin–destination (OD) matrices by identifying significant user locations using an individual-based regularity metric. Additionally, this study proposes an innovative route assignment approach that integrates CDR-derived trip patterns with the STRADA network model through a user-equilibrium traffic assignment process.

Applied to data from Sri Lanka’s Western Province and validated against a large-scale Household Visit Survey (HVS), the model demonstrates strong alignment with traditional travel demand estimates. The results affirm the potential of repurposed mobile phone

data—originally collected for billing purposes—as a reliable input for transport planning, matching or even surpassing traditional four-step modeling in terms of behavioral representation and scalability.

However, several limitations remain. The current methodology may misclassify users whose work and home locations fall within the same tower area because of spatial resolution constraints. Similarly, intra-cell movements are not captured, leading to potential underestimation of short trips. Moreover, the model assumes modal split proportions based on external data, as CDRs alone lack mode-specific indicators. Future research should explore the integration of high-frequency or multimodal datasets to enhance mode inference and capture within-cell trip variability, particularly for improving the modeling of public transport usage.

Nonetheless, it is important to note that the accuracy and generalizability of CDR-based models depend heavily on the penetration rate of mobile phone users and the density of tower infrastructure. In regions with limited mobile coverage or lower adoption rates—often rural or economically disadvantaged areas—the representativeness of the CDR sample may be compromised. This can lead to spatial data sparsity and underrepresentation of specific user groups, potentially limiting the scalability of this method in such contexts. Future extensions could incorporate multiple data sources or synthetic population weighting to mitigate these limitations.

This work represents a significant step toward the mainstream use of big mobile data in transport demand forecasting, offering a viable, scalable, and timely alternative to traditional methods. Additionally, while the CDR data used in this study is pseudonymized, we acknowledge the importance of ethical considerations in handling such data. Even anonymized records can carry re-identification risks if improperly processed. To address this, strict data access protocols were followed. Ethical data stewardship remains central to deploying big data solutions in transport modeling.

Author Contributions: Conceptualization, N.K.B.J. and A.S.K.; methodology, N.K.B.J.; software, N.K.B.J.; validation, N.K.B.J.; formal analysis, N.K.B.J.; investigation, N.K.B.J.; resources, N.K.B.J.; data curation, N.K.B.J.; writing—original draft preparation, N.K.B.J. writing—review and editing, N.K.B.J.; visualization, N.K.B.J.; supervision, A.S.K.; project administration, A.S.K.; funding acquisition, A.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to analysis of secondary anonymized data.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Lirneasia and are available from the authors with the permission of Lirneasia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Laharotte, P.-A.; Richard, W. Combining Usual Roadside Survey and Cellular Phone Data to Produce Accurate Transit and Exchange OD Matrices: Application to the Cross-Border Geneva 4-Step Model. *Transp. Res. Procedia* **2019**, *37*, 63–70. [[CrossRef](#)]
2. De Dios Ortúzar, J.; Willumsen, L.G. *Modelling Transport*, 4th ed.; Wiley: Chichester, UK, 2011.
3. Meyer, M.D. *Transportation Planning Handbook*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2016. [[CrossRef](#)]
4. Goulding, J. *Best Practices and Methodology for OD Matrix Creation from CDR Data*; N-LAB, University of Nottingham: Nottingham, UK, 2014.
5. Wang, M.; Schrock, S.D.; Vander, N.; Mulinazzi, T. Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data. *J. Urban Plan. Dev.* **2013**, *139*, 76–86. [[CrossRef](#)]
6. Gundlegård, D.; Rydergren, C.; Breyer, N.; Rajna, B. Travel Demand Estimation and Network Assignment Based on Cellular Network Data. *Comput. Commun.* **2016**, *95*, 29–42. [[CrossRef](#)]

7. Jiang, S.; Ferreira, J.; González, M.C. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Trans. Big Data* **2016**, *3*, 208–219. [[CrossRef](#)]
8. Maldeniya, D.; Kumarage, A.; Lokanathan, S.; Kreindler, G.; Madhawa, K. *Where Did You Come from? Where Did You Go? Robust Policy-Relevant Evidence from Mobile Network Big Data*; LIRNEAsia Working Paper; LIRNEAsia: Colombo, Sri Lanka, 2015; pp. 1–17.
9. Khan, F.H.; Ali, M.E.; Dev, H. Clustering and Association Rule Mining Based Traffic Analysis and Prediction of Dhaka. In Proceedings of the International Conference on Networking Systems and Security (NSysS), Dhaka, Bangladesh, 5–7 January 2015; pp. 1–6. Available online: <https://www.academia.edu/99960793/> (accessed on 16 May 2020).
10. Kung, K.S.; Greco, K.; Sobolevsky, S.; Ratti, C. Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data. *PLoS ONE* **2014**, *9*, e96180. [[CrossRef](#)] [[PubMed](#)]
11. Vieira, M.R.; Frías-Martínez, E.; Bakalov, P.; Frías-Martínez, V.; Tsostras, V.J. Querying Spatio-Temporal Patterns in Mobile Phone-Call Databases. In Proceedings of the IEEE International Conference on Mobile Data Management (MDM), Luleå, Sweden, 6–9 June 2010; pp. 239–248. [[CrossRef](#)]
12. Yang, P.; Zhu, T.; Wan, X.; Wang, X. Identifying Significant Places Using Multi-Day Call Detail Records. In Proceedings of the 26th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Limassol, Cyprus, 10–12 November 2014; pp. 360–366. [[CrossRef](#)]
13. Kung, K.S.; Sobolevsky, S.; Cottrill, C.D.; Bettencourt, L.M.A.; González, M.C.; Ratti, C. Understanding Aggregate Human Mobility Patterns Using Passive Mobile Phone Location Data—A Home-Based Approach. *Transportation* **2014**, *41*, 591–607. [[CrossRef](#)]
14. Fekih, M.; Bellemans, T.; Smoreda, Z.; Bonnel, P. A Data-Driven Approach of Origin-Destination Matrix Construction from Cellular Network Signalling Data: A Case Study of Lyon Region (France). *Transportation* **2020**, *47*, 2007–2036. [[CrossRef](#)]
15. Mamei, M.; Bicocchi, N.; Lippi, M.; Mariani, S.; Zambonelli, F. Evaluating Origin–Destination Matrices Obtained from CDR Data. *Sensors* **2019**, *19*, 4470. [[CrossRef](#)] [[PubMed](#)]
16. Bwambale, A.; Choudhury, C.F.; Hess, S. Modelling Trip Generation Using Mobile Phone Data: A Latent Demographics Approach. *J. Transp. Geogr.* **2017**, *58*, 19–29. [[CrossRef](#)]
17. Iqbal, S.; Choudhury, C.F.; Wang, P.; González, M.C. Development of Origin–Destination Matrices Using Mobile Phone Call Data. *Transp. Res. Part C Emerg. Technol.* **2014**, *40*, 63–74. [[CrossRef](#)]
18. Luo, X.; Zhou, Y.; Yang, Y.; Wu, S. Research on Home and Work Locations Based on Mobile Phone Data. *J. Phys. Conf. Ser.* **2020**, *1486*, 052013. [[CrossRef](#)]
19. Wang, F.; Chen, C. Mobility Analysis Workflow (MAW): An Accessible, Interoperable, and Reproducible Container System for Processing Raw Mobile Data. *Transp. Res. Part C Emerg. Technol.* **2023**, *148*, 103999.
20. Zagatti, G.A.; Chien, S.; Olken, B.A.; Wang, S.W.; Zaitchik, B.; Aberra, A. A Trip to Work: Estimation of Origin and Destination of Commuting Patterns in the Main Metropolitan Regions of Haiti Using CDR. *Dev. Eng.* **2018**, *3*, 133–165. [[CrossRef](#)]
21. Mamei, M.; Colonna, M.; Galassi, M. Automatic Identification of Relevant Places from Cellular Network Data. *Pervasive Mob. Comput.* **2016**, *31*, 147–158. [[CrossRef](#)]
22. Leng, Y.; Zhao, J.; Koutsopoulos, H.N. Leveraging Individual and Collective Regularity to Profile and Segment User Locations from Mobile Phone Data. *ACM Trans. Manag. Inf. Syst.* **2021**, *12*, 147–158. [[CrossRef](#)]
23. Lwin, K.K.; Sekimoto, Y.; Takeuchi, W. Estimation of Hourly Link Population and Flow Directions from Mobile CDR. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 449. [[CrossRef](#)]
24. Sakamane, P.; Phithakkitnukoon, S.; Smoreda, Z.; Ratti, C. Methods for Inferring Route Choice of Commuting Trip from Mobile Phone Network Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 306. [[CrossRef](#)]
25. Breyer, N.; Gundlegård, D.; Rydergren, C. Cellpath Routing and Route Traffic Flow Estimation Based on Cellular Network Data. *J. Urban Technol.* **2018**, *25*, 85–104. [[CrossRef](#)]
26. Japan International Cooperation Agency (JICA); Oriental Consultants Co., Ltd. *Urban Transport System Development Project for Colombo Metropolitan Region and Suburbs*; JICA: Tokyo, Japan, 2014; Available online: <https://openjicareport.jica.go.jp/pdf/12176665.pdf> (accessed on 17 May 2020).
27. Jeewanthi, N.K.B.; Kumarage, A.S. Home-Based Trip Estimation from Mobile Phone Data. In Proceedings of the 13th International Conference of the Eastern Asia Society for Transportation Studies (EASTS), Colombo, Sri Lanka, 9–12 September 2019; Available online: https://east.info/on-line/proceedings/vol.13/pdf/PP2918_R1_F.pdf (accessed on 10 January 2021).
28. Ayesha, B.; Jeewanthi, B.; Chitraranjan, C.; Perera, A.S.; Kumarage, A.S. User Localization Based on Call Detail Record. In *Lecture Notes in Computer Science, Proceedings of the 21st International Conference on Human–Computer Interaction (HCII), Orlando, FL, USA, 26–31 July 2019*; Springer: Cham, Switzerland, 2019; Volume 11871, pp. 558–575. [[CrossRef](#)]
29. Järv, O.; Ahas, R.; Witlox, F. Understanding Monthly Variability in Human Activity Spaces: A Twelve-Month Study Using Mobile Phone Call Detail Records. *Transp. Res. Part C Emerg. Technol.* **2014**, *38*, 122–135. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.