



OPEN Detecting heavy trucks from mobile phone trajectories using image-based behavioral representations and deep learning models

Franco Basso^{1,2}✉, Félix des Rotours³, Tomás Maldonado⁴, Raúl Pezoa⁴ & Mauricio Varas⁵

This paper proposes an innovative methodology for detecting heavy trucks utilizing mobile phone data, addressing significant limitations inherent in traditional tracking methods, often characterized by high costs, intrusiveness, and incomplete data capture. By employing Call Detail Records (CDR) and introducing an image-inspired architecture, the study uses Convolutional Neural Networks (CNN) to model the microscopic behavioral patterns of mobile devices. Our numerical results show that our proposed approach outperforms more classical machine learning methods that rely only on aggregated features. This novel approach offers a scalable and cost-effective alternative to conventional methods, representing a pioneering application of image-based analytical techniques to mobile phone data within freight transport research. This work provides a robust tool for analyzing freight transport patterns, thereby supporting the development of strategies to mitigate the negative externalities of freight transportation while preserving its economic benefits.

Keywords Deep learning, Logistics, Call detail records

The improvements in transportation infrastructure and connectivity in the last decades have led to a rapid growth in intercity freight flow worldwide^{1,2}. While essential to the functioning of cities by producing high economic benefits^{3,4}, freight transport has its share of negative externalities. For example, freight transportation is one of the main drivers of the increasing energy consumption in the transportation sector^{5,6}, accounting for about 45% of total transport energy consumption. Consequently, the literature has recognized the importance of understanding goods movement patterns as a basis for developing strategies that mitigate these adverse impacts while preserving economic advantages^{3,7}.

Despite its importance, existing methods for tracking heavy trucks often involve high costs and intrusive technologies, making their widespread application challenging. Indeed, most prior work in this field has primarily focused on traditional tracking methods such as GPS-based systems. While these methods are effective and precise, they suffer from biases and incompleteness, especially in a highly fragmented industry such as freight transport. Recent contributions have explored several data sources for vehicle detection, including satellite imagery and toll data. However, these approaches face cost, scalability, and data precision limitations.

This paper presents a novel methodology for detecting heavy trucks using mobile phone trajectories derived from Call Detail Records (CDR). CDR data is continuously generated through user interactions with mobile devices, capturing the user's approximate location and the corresponding timestamp each time the network is accessed. To the best of our knowledge, the only prior work that has used CDR data to study freight transport is⁸, which focused on identifying last-mile delivery vehicles through classical machine learning techniques and standard freight variables. In contrast, the present study addresses the detection of heavy freight vehicles with a new methodological framework. Specifically, we use CDR data and feature engineering to construct an image-inspired architecture that captures the microscopic behavioral patterns of individual devices. This design enables

¹School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile. ²Instituto Sistemas Complejos de Ingeniería, Santiago, Chile. ³Ecole Polytechnique, Paris, France. ⁴Department of Industrial Engineering, Universidad Diego Portales, Santiago, Chile. ⁵Centro de Investigación en Sustentabilidad y Gestión Estratégica de Recursos, Facultad de Ingeniería, Universidad del Desarrollo, Santiago, Chile. ✉email: francobasso@gmail.com

the application of image processing techniques, such as Convolutional Neural Networks (CNNs), to identify devices associated with heavy truck drivers.

This paper's contribution is twofold. First, we propose a novel methodology that models mobile phone trajectory data as spatial heatmaps, transforming mobility behaviors into a visual format. Second, we evaluate whether incorporating these heatmap representations improves the detection of heavy trucks compared to using only numerical aggregated features. To this end, we adopt a standard CNN architecture widely recognized for its effectiveness in image-processing tasks. Importantly, our objective is not to benchmark various deep learning models but rather to assess the marginal predictive improvement provided by heatmap-based representations. To our knowledge, this is the first effort in the freight transport literature to introduce an image-based deep learning framework over mobile phone data, demonstrating significant gains in detection accuracy.

The rest of the article is structured as follows: In Section 2, we review the literature related to the use of CDR data to study the mobility of people and goods. In Section 3, we describe the data, whereas in Section 4, we present the proposed methodology comprehensively. In Section 5, we apply all the presented methods. Finally, in Section 6, we present some concluding remarks and provide some avenues for future research.

Literature review

Deep learning and image-based models in transportation research

The application of deep learning in transportation research has grown rapidly, offering new tools to model the complex, nonlinear dynamics of mobility systems^{9,10}. Traditional statistical approaches like linear regressions or decision trees often fall short when dealing with the intricate spatiotemporal patterns of traffic flows¹¹, demand surges¹², and crash risks¹³. Deep architectures, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated superior predictive power by taking advantage of rich transportation datasets, including automatic vehicle identification¹⁴, mobile phone traces¹⁵, and connected vehicle data¹⁶.

A particularly promising innovation has been the use of image-based encodings to represent transportation data¹⁷. By transforming traffic states, spatial flows, or historical sequences into grid-based images, researchers enable CNNs—originally designed for visual tasks—to learn complex mobility patterns more effectively. This approach captures spatial correlations and temporal dynamics simultaneously, and has been applied successfully in traffic data imputation¹⁸, traffic speed prediction¹⁹, crash risk assessment²⁰, and truck activity identification²¹. Moreover, image-based representations facilitate the integration of heterogeneous sources²², such as weather maps²³, road layouts²⁴, and vehicle density heatmaps²⁵, into a unified predictive framework.

While deep models automatically learn from raw data, careful feature engineering remains crucial to fully use image-based inputs and other structured information²⁶. Preprocessing steps such as spatial aggregation, lag feature construction, and contextual variable encoding (e.g., weather conditions, holidays) significantly enhance model performance²⁷. In fact, the process of structuring transportation data into image-like grids itself acts as a form of spatial feature engineering, organizing raw inputs to maximize the learning capabilities of CNNs and related architectures²⁸. Such strategies help address common transportation data challenges, including sparsity²⁹, noise³⁰, and missing information²⁶, improving the robustness and reliability of predictive models across different contexts.

The use of CDR in transportation and mobility

In recent years, CDR has become a critical data source in transportation research³¹. Particularly, CDR allows the exploration of mobility patterns using data generated from widespread mobile phone use, including details on location, time, and duration of calls³². The adoption of CDR offers a cost-effective alternative to traditional data-gathering techniques³³, supporting the analysis of large-scale population movements³⁴ and urban dense areas identification³⁵, while also facilitating improvements in transportation network modeling³⁶.

The use of CDR poses challenges in accuracy compared to GPS or other tracking methods due to its spatial granularity and uncertainties since it was initially intended for billing and network management³⁷, not for tracking vehicular movements with high precision³⁸. Indeed, previous literature emphasizes the importance of correctly interpreting the data to avoid misplacement of devices within coverage areas³⁹. Consequently, using CDR data requires careful interpretation⁴⁰.

A well-established research line has focused on using CDR data to compute origin-destination (OD) matrices⁴¹, for example, develops a methodology that combines mobile phone CDR and limited traffic counts from some locations in Dhaka, Bangladesh, to generate and scale OD patterns over different periods. Similarly,⁴² outlines a procedure to extract average daily origin-destination trips by purpose and time of day and apply this to CDRs gathered from Boston, USA. In South America,³² proposes an algorithm combining CDR and sociodemographic information to estimate OD matrices in Rio de Janeiro, Brazil. All these works support the idea that CDR is a robust and cost-effective urban planning and monitoring tool.

In another research stream, mobile phone data was also used to determine travel modes. We refer the reader to⁴³ for a comprehensive review. The first efforts in this venue used Global System for Mobile Communications (GSM) technologies⁴⁴. These contributions focused on identifying three main situations for individuals: stationary, walking, or driving, although, for the driving category, it was not possible to differentiate between private and public motorized vehicles. This issue has been tackled in more recent efforts.⁴⁵, for example, estimates the number of people that use public transport daily, which is particularly relevant for cities that do not count with an automatic fare collection system. On the other hand,⁴⁶ proposes a new methodology that, building on powerful machine learning models, allows for differentiating more transport modes, including tram, train, and bike. As⁴⁷ points out, this could help obtain periodic travel mode partition without needing costly and time-consuming surveys.

Variable	Description
IMSI	Anonymized identifier of the device issuing the connection request
Timestamp	Full date and time of the network interaction
Coordinates	Triangulated location of the device as latitude and longitude

Table 1. Structure of the CDR dataset.

Hashed IMSI	Timestamp	Coordinates
a3f9c2e4b8d6492a	2021-08-02 16:58:23	POINT (-70.607923 -33.444934)
e41b7a51dc4f4fdd	2021-08-02 14:25:14	POINT (-70.650298 -33.454597)
f5a8dcf0e2db4ce4	2021-08-02 18:30:44	POINT (-70.793916 -33.391698)
d02b9eae2f1c6d2	2021-08-02 17:58:33	POINT (-70.633739 -33.435391)
c4b57a63d1e945e7	2021-08-02 13:09:58	POINT (-70.560828 -33.420019)

Table 2. Sample of cleaned CDR data used in the analysis.

Finally, another research venue involves using CDR data to characterize tourism mobility. In this regard,⁴⁸ describes a framework that identifies tourists among the total population by analyzing CDR data and using several predefined rules.⁴⁹ proposes a methodological framework for generating national tourism statistics using mobile positioning data. The authors use data from Estonia’s two largest mobile phone operators, finding strong correlations with Estonian inbound statistics on accommodation.⁵⁰ combines CDR and Twitter data to predict tourism flows using graph neural networks in Spain.

The research gap

Although CDR data has been widely used to study human mobility, its application to freight transportation remains limited. To date,⁸ is the only study using CDR to identify freight vehicles, focusing on last-mile deliveries with traditional machine learning models. Moreover, image-based frameworks in the freight transportation literature are almost absent. This paper addresses these gaps by introducing a deep learning methodology that combines spatial heatmaps and aggregated features to detect heavy trucks, offering a scalable and robust alternative for freight transport analysis.

Data processing

In this paper, we deploy deep learning tools over cell phone data to detect heavy freight vehicles. Our research hypothesis is that the behavior of freight trucks can be synthesized from the connection data of an onboard device such as a cell phone. Since deep learning algorithms require a substantial volume of data to train effectively, a large amount of connection data is required. In what follows, we describe the data gathered from a mobile phone company we worked with, detailing the data cleaning process to mitigate problems related to CDR data, such as the ping-pong effect.

Raw data

We use data gathered from Empresa Nacional de Telecomunicaciones (Entel), one of the largest mobile operators in Chile. Entel offers several other services in addition to mobile telephony, including fixed telephony, Internet, and television. By 2021, the company had about a third of the wireless subscribers in Chile. While providing Internet access to its subscribers, Entel collects data for each connection named CDR.

The International Telecommunication Union (ITU), a UN agency founded in 1865, facilitates global communication by allocating radio frequencies, setting technical standards, and improving access to ICTs for underserved communities. ITU recommendations, like E.212, which defines the 15-character International Mobile Subscriber Identity (IMSI), are widely respected. Following E.212, CDR data is anonymized, with IMSI representing each user. For each network connection, the device from which the request originates simultaneously calls several antennas - a base antenna supported by two or three support antennas, depending on the quality of coverage. During this process, the network antennas identify the IMSI, triangulate its position, and record the connection time. Table 1 describes the structure of the CDR dataset. Table 2 presents a sample of the cleaned CDR data used in the study, showing anonymized IMSI identifiers, timestamps, and triangulated location coordinates.

CDRs are automatically collected by mobile operators to bill for connections. This collection mode is more economical and more accessible to repeat than mobility surveys carried out by statistical institutes or public players. Furthermore, the granularity of CDR data gives tracking a far greater spatial precision and temporal spread than surveys, which trace household movements over an arbitrary day. They are also immune to moral hazard, whereas the validity of surveys relies on the word of individuals who may lie or omit elements. Despite

these advantages, cell phone data lacks the information generally available in travel surveys concerning a respondent (e.g., age or income) or the conditions of his or her trip (e.g., purpose or mode). In the case of the CDRs we had available for this study, only daily data were available from 7 am to midnight. This limitation implies gaps in the trajectories, especially as much truck traffic occurs at night. What's more, the spatial triangulation performed by the antennas provides less spatial precision than the GPS method and sometimes even erroneous information.

Data cleaning

Preprocessing is necessary to clean the raw mobile phone data and reduce noise in the recorded trajectories. This section outlines the main steps we followed to prepare the dataset.

Correction of antenna-related location noise. Mobile operators often redistribute connections across nearby antennas to manage network load. This can result in sudden location changes, even when the user remains still, a phenomenon known as the *load-sharing effect*. Another issue is the *ping-pong effect*, where a device near the boundary between two antennas is alternately assigned to each one, producing artificial movement between two nearby points. To address this, we remove intermediate points when the angle formed between three consecutive positions falls within $[-5^\circ, 5^\circ]$ and the speed between them exceeds 120 km/h. Figures 1a and 1b show how this filtering works in practice.

Removal of machine-generated IMSIs. Some IMSIs in the dataset come from machines rather than people, such as payment terminals or surveillance systems. These machine-to-machine (M2M) devices can be identified using a registry provided by Entel, which includes 470,619 known M2M IMSIs in Chile. We exclude these from the analysis.

Filtering of high-frequency anomalies. Certain IMSIs generate over 10,000 records per day, which is far beyond normal usage patterns. These are treated as anomalies and removed from the dataset.

Averaging coordinates over time. To reduce noise and smooth the trajectories, we average the coordinates of each IMSI per minute. This helps remove erratic short-term variations without losing relevant movement patterns.

Methodology

Stays and trips

Since the seminal contribution of⁵¹, the concepts of *stays*, *destinations*, and *trips* have become well-known for understanding the location history of tracking devices.

In our context, a *stay* is defined as a single instance of a device spending some time in one place. A stay is detected when successive location points of a device remain within a radius of D meters for more than a duration of T seconds. The stay is then characterized by the medoid of its points. A *trip* is simply defined as the segment between two stays of the same device. Figure 2a shows three different stays and two trips, with the red dots representing the medoids of the respective stays.

In contrast, a *destination* is defined as any place where one or more devices have experienced a stay. In other words, a destination can be thought of as a timeless counterpart of stays. Note, however, that two close stays can refer to the same destination (for example, a logistic facility). Thus, a clustering algorithm is necessary to group close stays into a single destination. In this paper, we use the DBSCAN algorithm⁵² for this purpose, as it allows us to specify the radius R of the clusters rather than the number of clusters, which in our context leads to easier interpretation. We then refine the definition of trips and define a *long trip* as a trip occurring between two different destinations. Figure 2b shows two different destinations and one long trip between them.

Classification method

Our classification method utilizes two primary input data sources. The first is based on images and corresponds to a heatmap of the IMSI's interactions. The second consists of aggregated features. Figure 3 presents a general

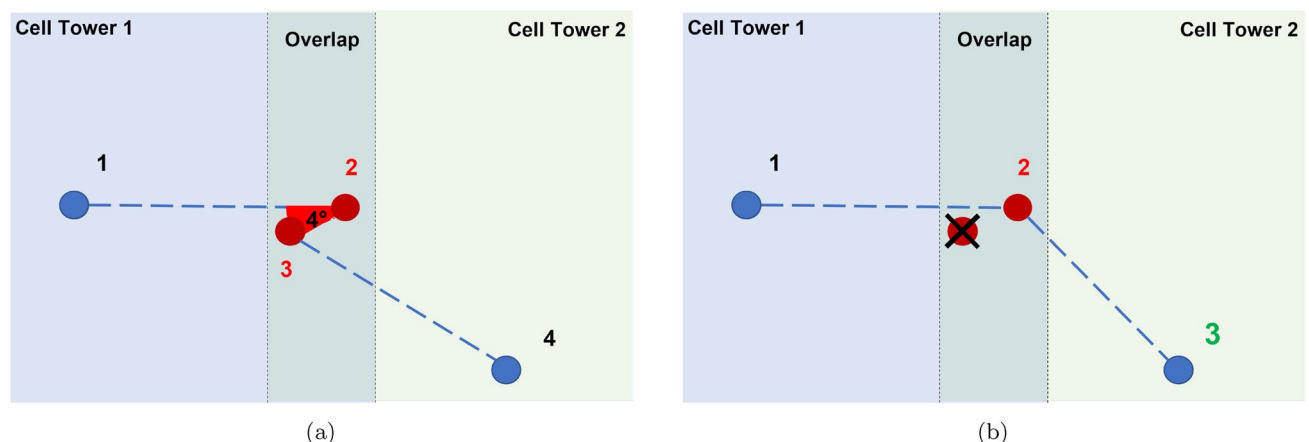


Fig. 1. Eliminating the ping pong effect.

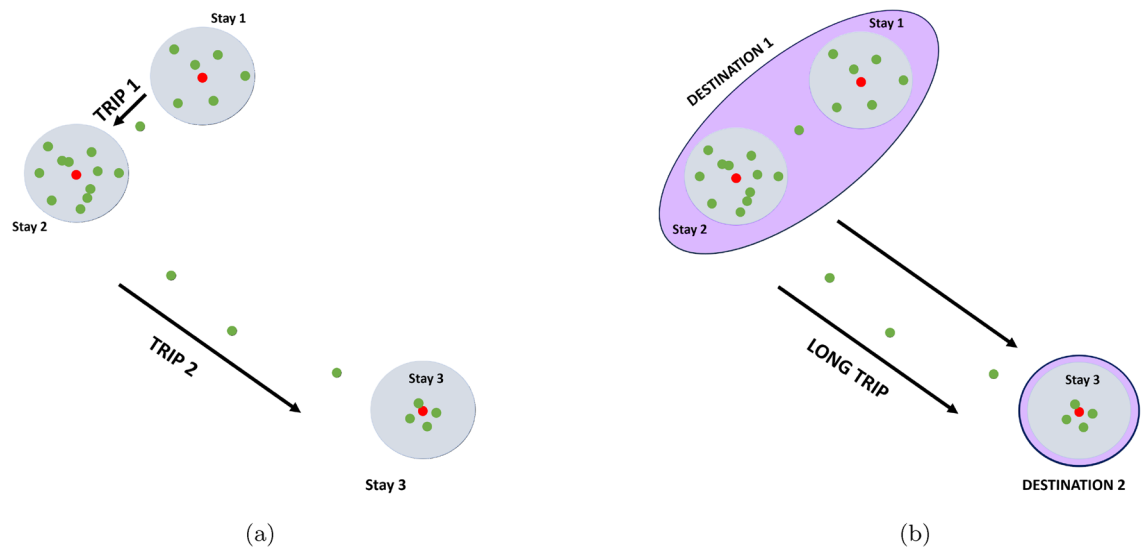


Fig. 2. Illustration of stays, trips, destinations and long trips.

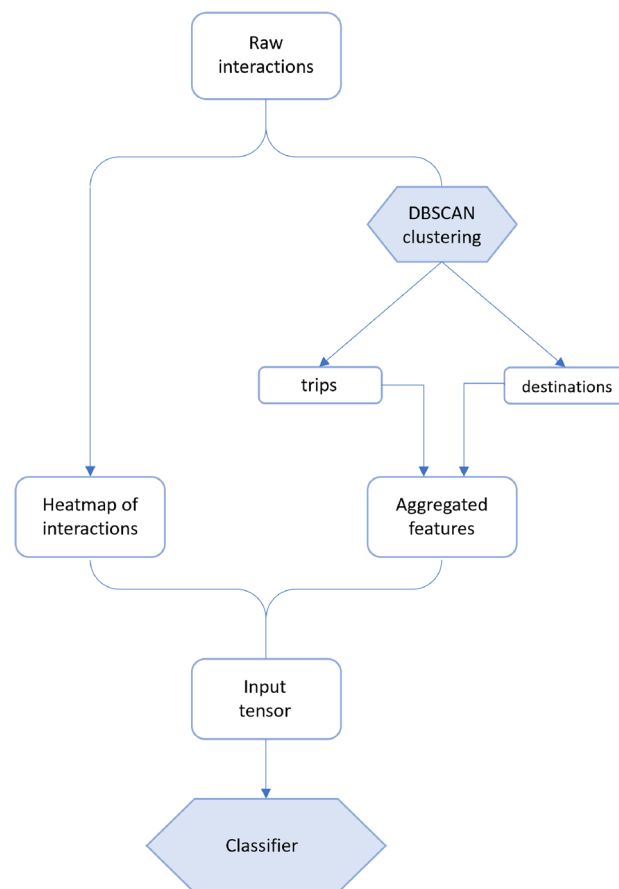


Fig. 3. Schematic diagram of the pipeline processing the data to feed the classifier.

framework of the proposed method, illustrating how the raw data is processed through parallel pipelines to feed a neural network.

On the right-hand side, interactions are labeled as either destinations or long trips using DBSCAN clustering, which allows for the computation of relevant aggregated features. These features are then concatenated with the heatmap of interactions from the left-hand pipeline. The resulting tensor is subsequently fed into the classifier.

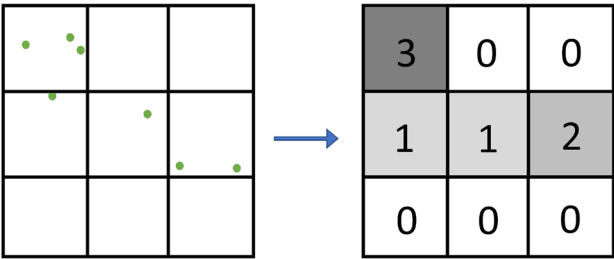


Fig. 4. Transformation of a raw interactions into a spatial heatmap.

Features	Feature importance
Destination number per day	0.155
Mean speed within long trips	0.129
Number of interactions within long trips per day	0.126
Mean duration within a destination	0.103
H3 number per day	0.100
Mean speed within destinations	0.083
Total interactions per day	0.080
Average cumulative duration of long trips per day	0.079
Average cumulative duration of destinations per day	0.073
Average number of interactions within destination per day	0.072

Table 3. Variable importance according to mean decrease Gini.

The following subsections look at data processing in each pipeline in more detail.

Heatmap of interactions

One of the main contributions of this research is the design of a new data architecture for the spatial representation of IMSI-related CDRs. Given a time interval—5 days in the case of the study—the interactions in a predetermined area of interest are grouped together in a heat map. This reduces the size of the data set fed into the neural network and improves the permanence zones of an IMSI. This approach is inspired on²⁰ accident prediction which proposes a bar code image-based data architecture for predicting crashes in a highway. Figure 4 illustrates the transformation of raw data into a heat map.

We define the spatial unit of the heatmap as the length of the side of each square cell. The resolution must be fine enough to capture meaningful variations in device density while avoiding excessive sparsity that could hinder model learning. After preliminary testing, the spatial resolution was set at 600 meters by 600 meters, balancing spatial detail and data density. Finer resolutions (e.g., 100 meters) resulted in highly sparse heatmaps, impairing the CNN’s ability to learn coherent spatial patterns. Conversely, coarser resolutions (e.g., 1 kilometer) smoothed out critical behavioral differences and diluted the distinctive spatial signatures of freight-related locations, such as ports and customs checkpoints. Thus, a 600-meter resolution was selected to ensure sufficient granularity to represent key freight activity patterns while maintaining computational efficiency and minimizing the risk of overfitting.

Aggregated features

Using a geospatial heatmap as input to the classifier is one of the main innovations in this study. However, while this input effectively captures spatial patterns, which is crucial for determining whether an IMSI is frequenting locations typical of a specific type of vehicle, it does not address the temporal dimension of movements. For instance, the trained algorithm may recognize a high-load truck passing through freight ports, road customs, or large warehouses. However, this criterion alone does not differentiate high-tonnage trucks from last-mile delivery trucks or even employees whose workplaces are at these locations. Therefore, it is necessary to introduce additional inputs that reflect the time dimension of journeys.

To address this, we adopt a similar approach to⁸, who used aggregated data to characterize IMSI behaviors and trained several classical classifiers. We consider several variables (see Table 3), such as the number of stays or destinations per day. To select the most relevant variables, we calibrate a random forest model and compute the Gini index, which helps determine the importance of each feature.

Convolutional neural network (CNN)

We use deep learning algorithms based on convolutional neural networks (CNNs) to generate predictions for each IMSI class. Introduced by⁵³, CNNs are designed for image recognition tasks, initially classifying handwritten digits to recognize visual patterns from pixel data. Unlike traditional neural networks, CNNs employ convolution

operations in at least one layer instead of general matrix products, reducing the number of fitted parameters⁵⁴. Typically, in CNNs, convolution is applied to an input image I using a two-dimensional kernel K , as shown in Eq. (1).

$$S(i, j) = (I * K)(i, j) = \sum_{m=0}^{r_i-1} \sum_{n=0}^{c_k-1} I(m, n) K(i-m, j-n) \quad (1)$$

where $S(i, j)$ represents the element of the output matrix at row i and column j , I is the input matrix, K is the kernel or filter, while r_i and c_k denote the number of rows and columns in the input matrix I and kernel K respectively.

A filter is then applied to the convolution output. This step reduces the dimension, for example, by calculating the local average of pixels (mean pooling) or the local maximum (max pooling) while highlighting certain image features. Several kernels can be applied in parallel to select distinct features. The error rate between predicted and true classes is calculated during each iteration to adjust the model parameters.

In order to take advantage from both heatmaps and aggregated features, we design a two-fold neural network:

- First features are extracted from the heatmap using a set of Convolutional and Max Pooling layer.
- Then a classifier labels the vector resulting from the concatenation of the extracted features and the aggregated features as a high load truck (1) or not (0)

Figure 5 illustrates this process. The red arrows represent the feedback loop, i.e. how the parameters of the different layers are adjusted according to the classification error at each epoch.

Figure 6 shows more precisely the different steps of data processing throughout the layers of the neural network.

It is important to point out that our objective is not to benchmark or exhaustively compare the performance of various deep learning architectures. Instead, we specifically aim to assess whether representing device behaviors as spatial heatmaps enhances the predictive accuracy of heavy truck detection compared to using only numerical aggregated features. We deliberately adopt a standard CNN architecture to isolate and clearly attribute any observed improvements to the heatmap representation itself. CNNs are widely recognized as the benchmark approach for processing image-based data, offering strong generalization capabilities with comparatively low model complexity. Given the structured nature of the heatmaps we generate—where spatial patterns are key to classification—using a classical CNN is sufficient to capture the relevant spatial dependencies without the added complexity of more advanced architectures. This modeling choice ensures that any performance improvement can be attributed to the informational richness of the heatmap representation rather than to incremental gains from using increasingly sophisticated deep learning models.

Training set

Our methodology aims to train a supervised machine learning model capable of determining whether an IMSI belongs to a heavy truck driver. To do so, however, we require a training dataset, with some IMSIs labeled as truck drivers, and others IMSIs labeled as non-truck drivers. However, note that our data is unlabeled. Consequently, we devise a procedure to build a training dataset.

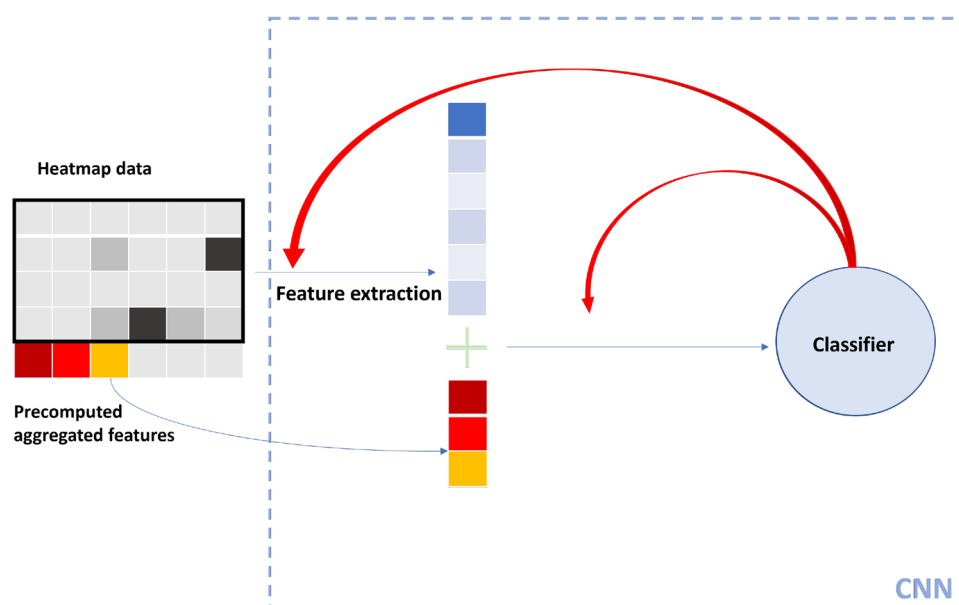


Fig. 5. Schematic diagram of the CNN.

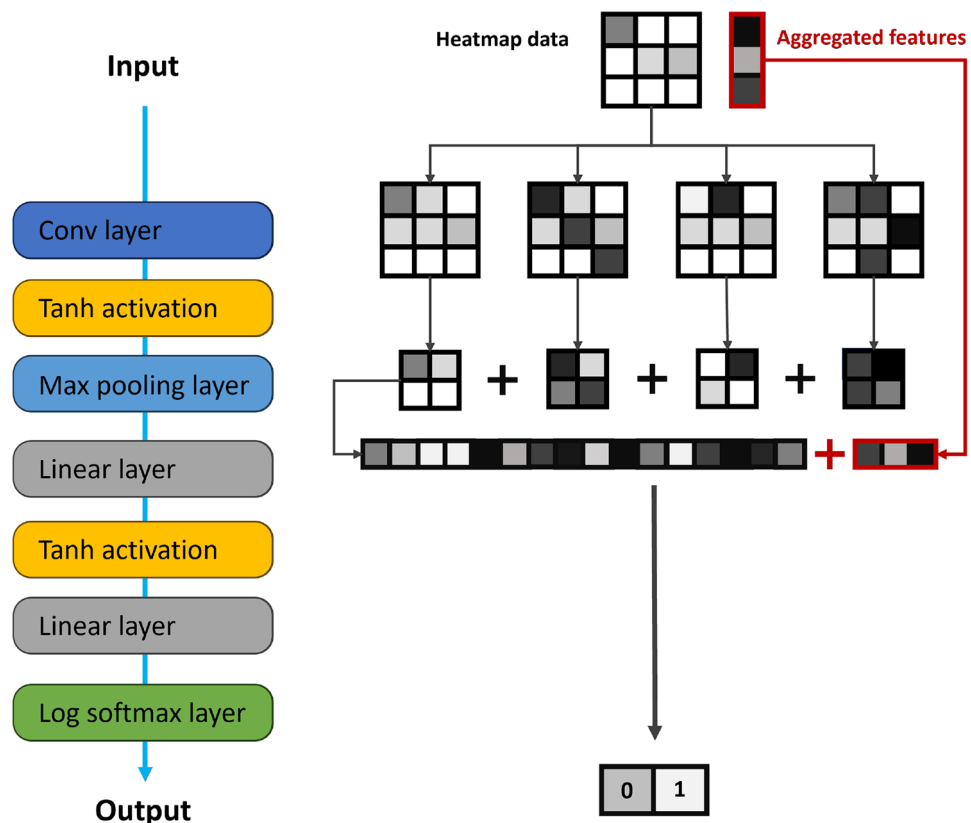


Fig. 6. Neural network architecture.

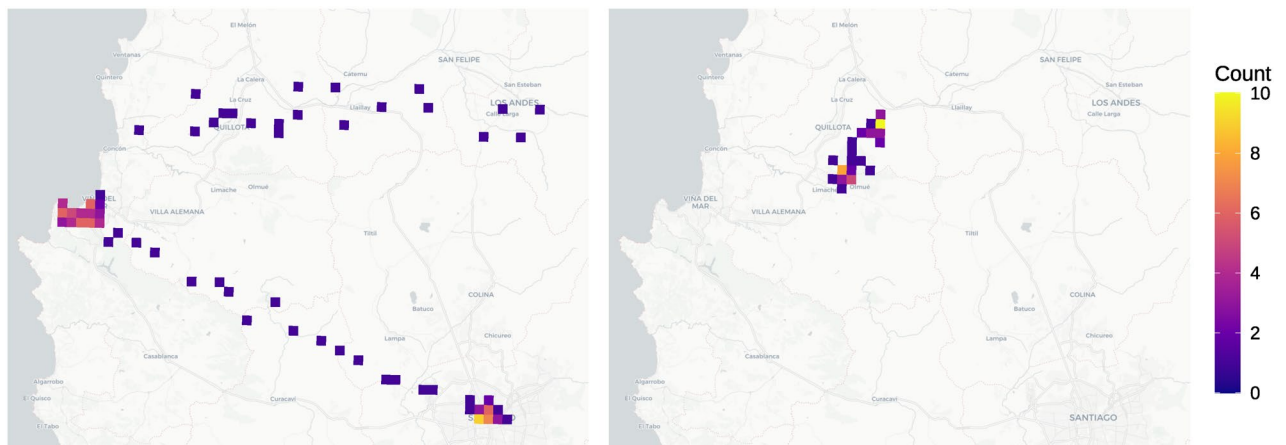


Fig. 7. Left: IMSI likely associated with a truck driver. Right: IMSI likely not associated with a truck driver.

Our approach begins by dividing the study area into grid cells and counting the number of interactions recorded in each square. The interaction counts are then visualized using a color scale proportional to the number of interactions per cell. Based on this representation, we manually identify a small subset of IMSIs that likely correspond to heavy truck drivers or non-truck users. Figure 7 provides an example of this process: the IMSI on the left shows interaction concentrations at key freight-related locations, such as Valparaíso's port and the customs facility near Argentina, while the IMSI on the right shows localized activity within smaller villages, which is unlikely to correspond to truck behavior.

We follow this manual procedure until we have selected n IMSIs similar to the first example and label them as heavy trucks (set A_0). Then, we select $k \cdot n$ random IMSIs from the rest and label them as non-trucks (set B_0). This approach is based on the assumption that random IMSIs are highly likely not to be linked to truck drivers. Subsequently, we calibrate a logistic regression over the set of IMSIs $C_0 = A_0 \cup B_0$ and apply it to the

remaining IMSIs to find candidates likely to be heavy trucks. These candidates are then visually inspected to decide whether they likely belong to heavy truck drivers. The intuition behind this procedure is that we cannot visually check every IMSI, so we use logistic regression to reduce the search space. We then complement this dataset with a random set of IMSIs (labeling them as non-trucks) to maintain the ratio 1 : k between trucks and non-trucks. We repeat this procedure until we have m IMSIs labeled as trucks.

Input: Set of IMSIs I . Integers n, m, k

```

1:  $i = 0$ 
2:  $I_0 = I$ 
3: Manually identify a subset  $A_0 \subseteq I_0$  of  $n$  IMSIs likely belonging to heavy truck drivers.
4: Select a random subset  $B_0 \subseteq I_0 \setminus A_0$  of  $k \cdot n$  IMSIs.
5: while  $|\cup_{j=0}^i A_j| \leq m$  do
6:   Build a calibration set  $C_i = \cup_{j=0}^i (A_j \cup B_j)$ , with a set of aggregated independent variables  $X_i$ ,
   and dependent variable  $Y_i = \mathbb{1}_{\cup_{j=0}^i A_j}$ .
7:   Build a test set  $T_i = I \setminus C_i$ , with a set of aggregated independent variables  $Z_i$ .
8:   Calibrate a logistic regression  $L_i$  over  $\{X_i, Y_i\}$ .
9:   Apply the logistic regression  $L_i$  over  $Z_i$ , identifying a set of heavy truck drivers candidates  $I_{i+1}$ .
10:  Manually identify the subset  $A_{i+1} \subseteq I_{i+1}$  that likely belong to heavy truck drivers.
11:  Select a random subset  $B_{i+1} \subseteq T_i \setminus I_{i+1}$  of  $k \cdot |A_{i+1}|$  IMSIs.
12:   $i = i + 1$ 
13: end while

```

Algorithm 1. Pseudocode for training dataset building.

Finally, the same visual representation shown in Fig. 7 serves as the basis for constructing the input fed to the CNN model. As shown in Fig. 8, the same spatial aggregation of interactions is retained, but the underlying basemap is removed and the interaction counts are normalized into a grayscale scale. These standardized heatmaps are then used as inputs to the CNN, allowing the model to learn spatial behavioral patterns.

Experimental results

The experimental design of this study follows an ablation framework aimed at isolating the individual contributions of different input representations to heavy truck detection. A similar experimental design has been carried out in^{55,56}, and⁵⁷. We define three experimental configurations: (i) classical machine learning models trained solely on aggregated numerical features derived from mobile phone trajectories, (ii) a CNN trained exclusively on spatial heatmaps capturing the behavioral patterns of devices, and (iii) a CNN trained on the fusion of heatmaps and aggregated features. This setup allows us to independently evaluate the predictive power of spatial information and traditional aggregated indicators, as well as their potential complementarity. By maintaining model complexity constant within each input type (e.g., using a standard CNN for image-based inputs), we ensure that any observed performance differences can be explicitly attributed to the nature of the data representation rather than to differences in modeling sophistication. This ablation framework thus enables a systematic and transparent assessment of the marginal value introduced by incorporating heatmap-based representations into freight vehicle detection models.

To implement the proposed methodologies, we apply the described procedures to a dataset of 17,000 IMSIs provided by Entel. These IMSIs were randomly selected under two conditions: each had at least one recorded



Fig. 8. Inputs to CNN model. Left: IMSI likely associated with a truck driver. Right: IMSI likely not associated with a truck driver.

interaction during the week of August 2nd to August 6th, 2021, and each was located within one of three key logistic regions in Chile—Metropolitan, Valparaíso, and O’Higgins. For this sample, we apply the data cleaning procedures detailed in Section 3.2, followed by the computation of stays, trips, long trips, and destinations as defined in Subsection 4.1. For the DBSCAN clustering algorithm used to identify destinations, we set the maximum distance between a point and its nearest neighbor to $\epsilon = 2700$ meters and the minimum number of points required to form a cluster to $\nu = 4$.

Interactions are grouped into rectangles measuring 600 meters by 600 meters. Each rectangle corresponds to one pixel in the heatmap, with intensity based on the number of interactions (as previously depicted in Figure 4). These dimensions strike a balance between reducing data size and maintaining spatial precision, making high-concentration zones more apparent. Ultimately, a 3D tensor of size [260, 277, 1] is generated for each IMSI. The first dimension corresponds to the x-axis of the heatmap, the second to the y-axis, and the third to the number of interactions in each spatial unit.

We consider ten aggregated features to cope with other non-spatial variables. Using a random forest, we compute the feature importance for these ten variables according to the mean decrease of the Gini’s impurity (Table 3). These computations require labeled data, for which we use the proposed method explained in Subsection 4.3, taking $n = 30$ for the IMSIs inspected manually, $m = 150$ for the IMSIs belonging to truck drivers, and a ratio value of $k = 2$, which leads to 300 non-truck IMSIs. Overall, our labeled dataset has 450 observations.

From Table 3, it follows that three most important features to detect truck are (i) number of destinations, (ii) mean speed, (iii) number of interactions. One hypothesis that explain the relevance of the number of destinations per day is the typical behavior of heavy trucks, which often make multiple stops at freight-related locations, such as warehouses, distribution centers, or ports. This pattern is less common for non-freight vehicles, making it a strong indicator of truck activity. Similarly, the mean speed is also a good indicator of truck activity: heavy trucks typically have different speed patterns compared to passenger vehicles due to their size, load, and the nature of their routes, often involving highways or industrial zones. Lower average speeds could indicate frequent stops or slower travel speeds, typical of trucks navigating urban areas or approaching loading/unloading sites. Finally, the number of interactions likely captures the intensity and frequency of stops, reflecting the operational patterns of trucks as they make deliveries or pickups.

We now compare the performance of the proposed approach with standard machine learning classification methodologies calibrated using only the aggregated feature dataset: logistic regression, SVM, and Random Forest. For this, we randomly select 75% of the labeled dataset for calibration and use the rest for validation. For the case of the CNN, we conducted a grid search procedure to tune the hyperparameters.

The final model was trained for 80 epochs using a learning rate of 0.001, the Adam optimizer, a hyperbolic tangent activation function, and cross-entropy loss. Table 4 summarizes the evaluation metrics, and Fig. 9 shows the ROC curves for the different models. Each ROC curve shows the trade-off between the true positive rate (recall) and the false positive rate for different classification thresholds. The Area Under the Curve (AUC) measures overall model performance, where a higher AUC means better discrimination between heavy truck and non-truck IMSIs.

As shown in Fig. 9, the CNN model that uses spatial trajectories achieves better performance than the classical machine learning models based on aggregated features. The concatenated CNN model, which combines both heatmaps and aggregated features, reaches the highest AUC. It also obtains better precision, accuracy, recall, and F1 score, showing that combining both sources of information improves truck detection. In addition, Fig. 9 shows that the concatenated CNN model maintains very high sensitivity even at very low false positive rates. This behavior is important in practical applications where the goal is to detect heavy trucks reliably while minimizing false positives, such as in freight monitoring or logistic management systems.

Concluding remarks

The efficient monitoring of freight transportation is crucial for urban logistics and economic activities, especially in rapidly growing urban areas⁸. This is particularly true for intercity freight transport, which has broader urban implications as it contributes to regional integration by redistributing economic activities². Traditional methods of tracking heavy trucks are often costly and lack scalability, limiting the decision-making support they provide. This study uses mobile phone CDR data within a novel deep learning framework to identify devices belonging to heavy truck drivers. Given the granularity and cost-effectiveness of CDR data, we believe our approach offers a high-potential solution for characterizing and studying freight systems.

Our proposed approach offers two main advantages. First, the graphical representation of devices’ behaviors captures disaggregated information that is lost when using simple aggregated values. Second, this graphical representation makes it natural to use state-of-the-art deep learning techniques designed for image processing.

Model	Dataset	Accuracy	Precision	Recall	F1 score	AUC ROC
Logistic regression	Aggregated	0.837	0.818	0.711	0.761	0.895
Random forest	Aggregated	0.827	0.833	0.658	0.735	0.854
SVM (kernel rbf)	Aggregated	0.817	0.732	0.789	0.759	0.853
CNN	Heatmap	0.897	0.877	0.833	0.856	0.9294
CNN (concatenated)	Aggregated/Heatmap	0.948	0.939	0.921	0.929	0.989

Table 4. Comparison of performance indicators for different models.

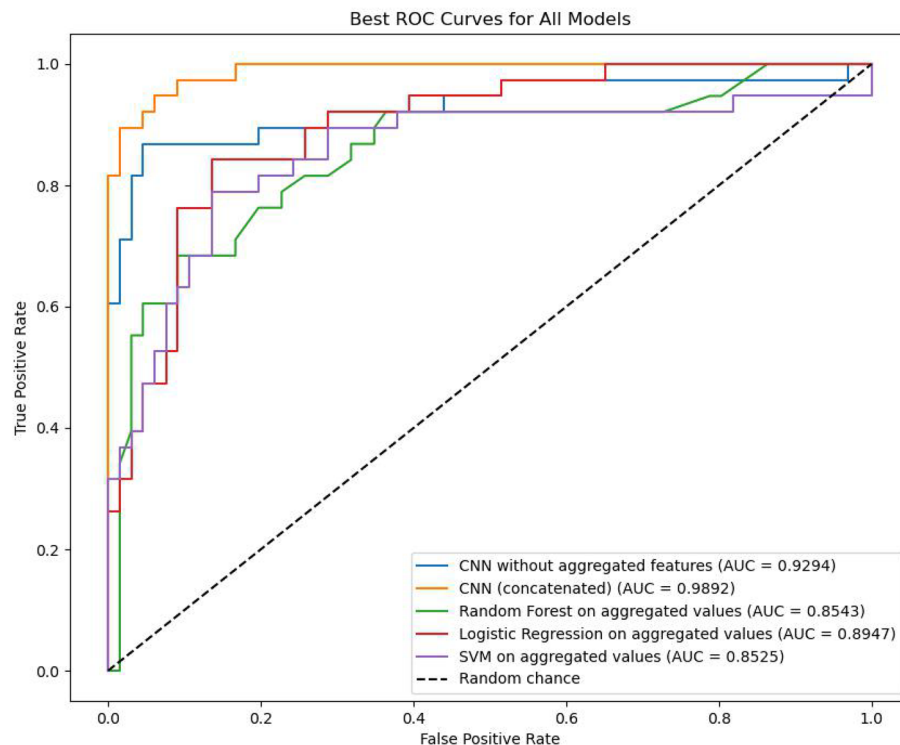


Fig. 9. ROC curves for the different models.

The main takeaway of this work is that our proposed architecture significantly outperforms traditional approaches based only on aggregated features. However, the best results are achieved when combining both type of features: image-based and aggregated data. Indeed, using only images reduces classifier performance because, while route images provide additional information, they lack the ability to represent key time-based attributes such as stay duration or mean speeds.

From this work, it follows that using deep learning algorithms for CDR data is promising for identifying and characterizing freight transportation, providing a robust support for decision-making. We believe the practical utility of our approach is clear: the combination of our adaptable methodology with a scalable technology such as CDR makes it suitable for large-scale logistics applications across various contexts and data volumes. Also, a byproduct of our approach is the graphical representation of the behavior of devices belonging to freight drivers. This output could help decision-makers to quickly identify, for example, congested zones, reacting consequently.

While CDR-based methods offer a scalable and cost-effective alternative for freight transport analysis, their value must be considered in light of the limitations associated with GPS-based approaches. In Chile, GPS usage is more common among large logistics firms but significantly lower among small and medium-sized carriers. This is especially critical given that the Chilean freight sector is highly fragmented, with many small operators. As a result, it is practically impossible to obtain comprehensive GPS data that represents the full spectrum of freight activity. Moreover, even when GPS systems are in place, access to this data is often restricted due to privacy concerns, proprietary ownership, and inconsistent spatial coverage. In contrast, CDR data enables observation at a much broader scale, allowing for estimating freight movement patterns across entire cities or even nationwide. This capability is particularly valuable for capturing behaviors in less digitized sector segments that would otherwise remain invisible to traditional monitoring systems.

Despite the promising results of our approach, several limitations must be acknowledged. First, our labeling strategy, based on manual inspection of IMSI trajectories and semi-automatic expansion using logistic regression, may introduce human biases. It could favor devices exhibiting *typical* heavy truck behaviors, such as regular visits to ports or logistics hubs, while underrepresenting atypical patterns like irregular or ad-hoc deliveries. Moreover, the labeled dataset (150 trucks versus 300 non-trucks) covers only a small portion of the 17,000 IMSIs analyzed, and the lack of exhaustive manual inspection raises the risk of false negatives, where truck drivers may have remained unlabeled. The absence of ground-truth validation, such as GPS-confirmed identifiers, also limits the assessment of labeling accuracy. Although multiple iterations of manual refinement were performed to diversify the training set, some residual biases may persist, particularly in classifying borderline mobility behaviors. Another limitation is that CDR data captures driver-owned mobile phones rather than telematics devices, causing personal and professional movements to overlap. Although we focused on daytime hours and incorporated features sensitive to freight patterns, full separation between personal and professional trips cannot be guaranteed.

Future research could address the limitations of this study by integrating external data sources, such as GPS records or fleet registries, to enhance label validation and coverage. Adopting semi-supervised or weakly

supervised learning frameworks could also offer more robust classification strategies under limited labeling conditions. Incorporating more refined temporal and spatial filters, as well as using information on freight-related locations, may further improve the separation between personal and freight mobility patterns, enhancing the reliability and applicability of CDR-based freight transport analysis. Beyond the scope of truck detection, future research could focus on expanding the applications of CDR data to other logistics topics, such as optimizing delivery routes, predicting traffic congestion, or improving fleet management. Investigating advanced machine learning models and integrating complementary data sources, including GPS and IoT sensors, could improve the accuracy and robustness of predictions. In addition, exploring real-time data processing techniques could enable dynamic decision-making, allowing rapid responses to changing traffic and demand conditions. Finally, future work could examine the impact of these advancements on cost savings, operational efficiency, and environmental sustainability, while collaboration with industry partners would facilitate the practical implementation of these methodologies to address real-world logistics challenges effectively. These directions would not only strengthen the methodological foundations of CDR-based freight analysis but also broaden its practical relevance for supporting data-driven decision-making in transportation logistics.

Data availability

The data that support the findings of this study are available from ENTEL but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors through the corresponding author, Franco Basso, upon reasonable request and with permission of ENTEL.

Received: 21 March 2025; Accepted: 10 June 2025

Published online: 01 July 2025

References

1. Liu, P., Dong, M. & Gong, D. Eliminating overload trucking via a modal shift to achieve intercity freight sustainability: A system dynamics approach. *Sustainability* **9**(3), 398 (2017).
2. Zhao, Y., Cheng, S., Liu, K., Zhang, B. & Feng, L. Intercity freight connections in China under the view of mass truck trajectories. *Cities* **150**, 105034 (2024).
3. Gonzalez-Calderon, C. A., Sánchez-Díaz, I., Sarmiento-Ordosgoitia, I. & Holguín-Veras, J. Characterization and analysis of metropolitan freight patterns in Medellín, Colombia. *Eur. Transport Res. Rev.* **10**, 1–11 (2018).
4. Strulak-Wójcikiewicz, R. & Wagner, N. Exploring opportunities of using the sharing economy in sustainable urban freight transport. *Sustain. Cities Soc.* **68**, 102778 (2021).
5. Guangtong, X., Lv, Y., Sun, H., Jianjun, W. & Yang, Z. Mobility and evaluation of intercity freight CO₂ emissions in an urban agglomeration. *Transport. Res. Part D Transport. Environ.* **91**, 102674 (2021).
6. Basso, F., Feijoo, F., Pezoa, R., Varas, M. & Vidal, B. The impact of electromobility in public transport: An estimation of energy consumption using disaggregated data in Santiago, Chile. *Energy* **286**, 129550 (2024).
7. Basso, F., Pezoa, R., Tapia, N. & Varas, M. Estimation of the origin-destination matrix for trucks that use highways: A case study in Chile. *Sustainability* **14**(5), 2645 (2022).
8. Basso, F., Cox, T., Pezoa, R., Maldonado, T. & Varas, M. Characterizing last-mile freight transportation using mobile phone data: The case of Santiago, Chile. *Transport. Res. Part A: Policy Pract.* **186**, 104149 (2024).
9. Haghighat, A. K. et al. Applications of deep learning in intelligent transportation systems. *J. Big Data Anal. Transport.* **2**, 115–145 (2020).
10. Abirami, S., Pethuraj, M., Uthayakumar, M. & Chitra, P. A systematic survey on big data and artificial intelligence algorithms for intelligent transportation system. *Case Stud. Transport Policy* 101247 (2024).
11. Yang, S., Ma, W., Pi, X. & Qian, S. A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources. *Transport. Res. Part C Emerg. Technol.* **107**, 248–265 (2019).
12. Mamede, F. P. et al. Deep learning and statistical models for forecasting transportation demand: A case study of multiple distribution centers. *Logistics* **7**(4), 86 (2023).
13. Theofilatos, A., Chen, C. & Antoniou, C. Comparing machine learning and deep learning methods for real-time crash prediction. *Transport. Res. Rec.* **2673**(8), 169–178 (2019).
14. Tang, K. et al. Dynamic origin-destination flow estimation using automatic vehicle identification data: A 3d convolutional neural network approach. *Comput.-Aided Civ. Infrastruct. Eng.* **36**(1), 30–46 (2021).
15. Liang, X., Zhang, Y., Wang, G. & Songhua, X. A deep learning model for transportation mode detection based on smartphone sensing data. *IEEE Trans. Intell. Transport. Syst.* **21**(12), 5223–5235 (2019).
16. Jiajie, H., Huang, M.-C. & Xiong, Y. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. *Accid. Anal. Prevent.* **144**, 105665 (2020).
17. Nguyen, H., Kieu, L.-M., Wen, T. & Cai, C. Deep learning methods in transportation domain: A review. *IET Intell. Transport Syst.* **12**(9), 998–1004 (2018).
18. Huang, T., Chakraborty, P. & Sharma, A. Deep convolutional generative adversarial networks for traffic data imputation encoding time series as images. *Int. J. Transport. Sci. Technol.* **12**(1), 1–18 (2023).
19. Ma, X. et al. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* **17**(4), 818 (2017).
20. Basso, F., Pezoa, R., Varas, M. & Villalobos, M. A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. *Accid. Anal. Prevent.* **162**, 106409 (2021).
21. Hu, J., Yang, S., Yang, Y., Xu, X. & Zhang, X. Identifying truck activity from global positioning system data based on satellite imagery. *Transport. Res. Rec.* 03611981241283012 (2024).
22. Bazzan, A. L. C. et al. Using information from heterogeneous sources and machine learning in intelligent transportation systems. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*. 213–220. (IEEE, 2019).
23. Yang, L., Muresan, R., Al-Dweik, A. & Hadjileontiadis, L. J. Image-based visibility estimation algorithm for intelligent transportation systems. *IEEE Access* **6**, 76728–76740 (2018).
24. Konstantinidis, F., Sackmann, M., Hofmann, U. & Stiller, C. Modeling interaction-aware driving behavior using graph-based representations and multi-agent reinforcement learning. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. 1643–1650. (IEEE, 2023).
25. Gao, H. et al. Identifying visual quality of rural road landscape character by using public preference and heatmap analysis in Sabak Bernam, Malaysia. *Land* **12**(7), 1440 (2023).

26. Zhang, Y., Shi, X., Zhang, S. & Abraham, A. A xgboost-based lane change prediction on time series data using feature engineering for autopilot vehicles. *IEEE Trans. Intell. Transport. Syst.* **23**(10), 19187–19200 (2022).
27. Lin, H., He, Y., Liu, Y., Gao, K. & Xiaobo, Q. Deep demand prediction: An enhanced conformer model with cold-start adaptation for origin-destination ride-hailing demand prediction. *IEEE Intell. Transport. Syst. Mag.* **16**(3), 111–124 (2023).
28. Essa, H. A., Ismaiel, E. & Hinnawi, M. F. A. Feature-based detection of breast cancer using convolutional neural network and feature engineering. *Sci. Rep.* **14**(1), 22215 (2024).
29. Sharma, S., Nayak, R. & Bhaskar, A. Multi-view feature engineering for day-to-day joint clustering of multiple traffic datasets. *Transport. Res. Part C: Emerg. Technol.* **162**, 104607 (2024).
30. Etamad, M., Soares Júnior, A. & Matwin, S. Predicting transportation modes of GPS trajectories using feature engineering and noise removal. In *Canadian Conference on Artificial Intelligence*. 259–264. (Springer, 2018).
31. Dypvik Landmark, A., Arnesen, P., Södersten, C.-J. & Hjelkrem, O.A. Mobile phone data in transportation research: Methods for benchmarking against other data sources. *Transportation*. 1–23. (2021).
32. Barboza, M. H. C. et al. Identifying human mobility patterns in the Rio de Janeiro metropolitan area using call detail records. *Transport. Res. Rec.* **2675**(4), 213–221 (2021).
33. Järvi, O., Ahas, R. & Witlox, F. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transport. Res. Part C Emerg. Technol.* **38**, 122–135 (2014).
34. Moral-Carcedo, J. Predicting traffic intensity in the urban area of Madrid: Integrating route network topology into a machine-learning model. *Eng. Appl. Artif. Intell.* **137**, 109154 (2024).
35. Rubio, A., Sanchez, A. & Frias-Martinez, E. Adaptive non-parametric identification of dense areas using cell phone records for urban analysis. *Eng. Appl. Artif. Intell.* **26**(1), 551–563 (2013).
36. Burkhard, O., Ahas, R., Saluveer, E. & Weibel, R. Extracting regular mobility patterns from sparse CDR data without a priori assumptions. *J. Loc. Based Serv.* **11**(2), 78–97 (2017).
37. Su, D. & Qi, F. An approach for ensuring the reliability of call detail records collection in billing system. In *2009 International Conference on Research Challenges in Computer Science*. 100–103. (IEEE, 2009).
38. Zhao, Z., Koutsopoulos, H. N. & Zhao, J. Identifying hidden visits from sparse call detail record data. *Trans. Urban Data Sci. Technol.* **1**(3–4), 121–141 (2022).
39. Jovanovic, V. M. & Cummings, B. T. Analysis of mobile phone geolocation methods used in US courts. *IEEE Access* **10**, 28037–28052 (2022).
40. Zhao, Z. et al. Understanding the bias of call detail records in human mobility research. *Int. J. Geogr. Inf. Sci.* **30**(9), 1738–1762 (2016).
41. Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. Development of origin–destination matrices using mobile phone call data. *Transport. Res. Part C Emerg. Technol.* **40**, 63–74 (2014).
42. Alexander, L., Jiang, S., Murga, M. & González, M. C. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transport. Res. Part C Emerg. Technol.* **58**, 240–250 (2015).
43. Huang, H., Cheng, Y. & Weibel, R. Transport mode detection based on mobile phone network data: A systematic review. *Transport. Res. Part C Emerg. Technol.* **101**, 297–312 (2019).
44. Sohn, T. et al. Mobility detection using everyday GSM traces. In *International Conference on Ubiquitous Computing*. 212–224. (Springer, 2006).
45. Vidović, K., Čolić, P., Vojvodić, S. & Blavicki, A. Methodology for public transport mode detection using telecom big data sets: Case study in Croatia. *Transport. Res. Proc.* **64**, 76–83 (2022).
46. Chin, K., Huang, H., Horn, C., Kasanicky, I. & Weibel, R. Inferring fine-grained transport modes from mobile phone cellular signaling data. *Comput. Environ. Urban Syst.* **77**, 101348 (2019).
47. Graells-Garrido, E., Caro, D. & Parra, D. Inferring modes of transportation using mobile phone data. *EPJ Data Sci.* **7**(1), 1–23 (2018).
48. Sikder, R., Uddin, M. J. & Halder, S. An efficient approach of identifying tourist by call detail record analysis. In *2016 International Workshop on Computational Intelligence (IWCI)*. 136–141. (IEEE, 2016).
49. Saluveer, E. et al. Methodological framework for producing national tourism statistics from mobile positioning data. *Ann. Tour. Res.* **81**, 102895 (2020).
50. Sáenz, F. T., Arcas-Tunéz, F. & Muñoz, A. Nation-wide touristic flow prediction with graph neural networks and heterogeneous open data. *Inf. Fusion* **91**, 582–597 (2023).
51. Hariharan, R. & Toyama, K. Project lachesis: Parsing and modeling location histories. In *International Conference on Geographic Information Science*. 106–124. (Springer, 2004).
52. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining* (1996).
53. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998).
54. Aghdam, H. & Heravi, E. *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*. Vol. 01 (2017).
55. Thakare, K. V., Dogra, D. P., Choi, H., Kim, H. & Kim, I.-J. Object interaction-based localization and description of road accident events using deep learning. *IEEE Trans. Intell. Transport. Syst.* **23**(11), 20601–20613 (2022).
56. Suha, S.A. & Islam, M.N. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Sci. Rep.* **12**(1), 17123 (2022).
57. Qikai, Q., Shen, Y., Yang, M., Zhang, R. & Zhang, H. Expressway traffic incident detection using a deep learning approach based on spatiotemporal features with multilevel fusion. *J. Transport. Eng. Part A Syst.* **150**(6), 04024020 (2024).

Acknowledgements

We acknowledge ENTEL for providing us with CDR data and cooperating enthusiastically with the project.

Author contributions

Franco Basso: Writing—review & editing, Writing—original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. Félix des Rotours: Writing—review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. Tomás Maldonado: Writing—review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Investigation, Formal analysis, Data curation. Raúl Pezoa: Writing—review & editing, Writing—original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. Mauricio Varas: Writing—review & editing, Writing—original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. All authors declare consent for publication.

Funding

Franco Basso gratefully acknowledges the financial support from the Complex Engineering Systems Institute, ISCI (ANID-PIA/PUENTE AFB230002), ANID-FONDECYT 1241335, ANID-FONDEF ID25I10409 and ANID-FOVI 240048. Raúl Pezoa acknowledges the financial support from ANID-FONDECYT 1241913. Mauricio Varas acknowledges the financial support from ANID-FONDECYT 1241843.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025