TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–12

DOI: 10.1177/03611981231158639
journals.sagepub.com/home/trr
SAGE

Freight Operational Characteristics Mined from Anonymous Mobile Sensor Data

Taslima Akter[1], Sarah Hernandez[2], and Pedro V. Camargo[3]

Abstract

Effective transportation performance measurement (TPM) benefits from ubiquitous transportation system monitoring both spatially and temporally. In the context of freight-oriented TPM, traditional devices such as inductive loops, cameras, manual counts, and so forth, may fail to provide comprehensive and high-resolution coverage, providing, for example, only volume counts for a small subset of links across a large network with no indication of trip linkages. New sources of big data from mobile sensors including on-board global positioning system (GPS) devices allow more comprehensive network coverage and insights into trip chaining behaviors. However, to gain actionable insights into system performance from large and noisy streams of mobile sensor data, it is necessary to mine it for relevant operational characteristics of the trucks it represents. Such characteristics include stop locations, stop duration, stop time of day, trip length, and trip duration. To address this methodological need, this paper presents three heuristic algorithms: "stop identification," "path identification," and "trip identification." To address the issue of determining relevant operational characteristics, a multinomial logit (MNL) model approach is applied to determine the commodity carried based on the outputs of the heuristic algorithms. The MNL model is novel in that it relates operational characteristics to commodity carried thus filling a critical data gap that currently limits the development of advanced freight forecasting models. The set of models developed in this paper allow large-scale GPS data to be used for freight planning while maintaining levels of data anonymity that allow such data to be shared with public agencies.

Keywords

data and data science, freight transportation data, GIS data and analysis, global positioning systems (GPS), model/modeling

Effective transportation performance measurement (TPM) benefits from ubiquitous system coverage. Because of the significant impact of trucking on the economy, infrastructure, and environment, it is essential that transportation agencies consider freight movements in their TPM strategies. In the U.S., to ensure freight needs are met, federal legislation (e.g., Fixing America's Surface Transportation [FAST] Act), mandates a process of selecting performance measures, setting performance targets, and establishing a freight plan that aligns with the broad goal of improving the national highway freight network to ensure economic competitiveness.

With the push toward more accurate and detailed freight performance measurement and system planning, there is an ever-increasing need to better understand and measure freight truck movements at high levels of temporal and spatial disaggregation (*1*). In the context of freight-oriented TPM, traditional performance monitoring devices such as inductive loops, cameras, manual counts, and so forth, may fail to provide comprehensive, high-resolution coverage of the transportation network. For instance, fixed location devices such as inductive loop detectors and cameras typically only provide volume and some vehicle classification data for the link on which they are located and give no indication of trip

[1]CPCS Transcom Inc., Washington, DC
[2]Department of Civil Engineering, University of Arkansas, Fayetteville, AR
[3]Outer Loop Consulting Pty, Brisbane, Australia

Corresponding Author:
Taslima Akter, takter@cpcstrans.com

linkages, for example, origin and destination, stop locations, and so forth. Acquiring the data needed for system-wide TPM is a challenge for transportation agencies and a special challenge of freight data when considering freight flows. Since freight operations are carried out primarily by private entities (e.g., shippers, carriers, and third-party logistics companies), operational data defining vehicle movements is often not made readily available for public consumption because of privacy concerns.

New sources of big data from mobile sensors including cell phones, electronic logging devices (ELDs) and global positioning system (GPS) devices allow more universal network coverage and, with that, the ability to gain insights into trip chaining behaviors. Carrier collectives have recently made available large streams of anonymized GPS data (*2*). This GPS data typically contains the timestamp, latitude and longitude position (e.g., ping), and point speed data for a sample of trucks operated by major freight carriers. All data about the carrier, fleet operator, driver, cargo/commodity, and trip purpose are removed from the data to protect privacy. Therefore, the anonymized data must be mined to extract relevant information for planning applications, such as stop location/purpose, trip purpose, and commodity carried, while maintaining the integrity of data-sharing agreements by ensuring the data remains anonymous. In other words, data mining should not reveal private information such as company/fleet identification or name.

Freight activity insights derived from truck GPS data have been applied in practice to support a variety of freight planning efforts including freight forecasting tools such as activity-based and truck touring models, estimating origin-destination truck flows, improving the estimation of freight performance measures, and ranking roadway bottlenecks (*3–10*). Although these studies used truck GPS data to develop and/or validate their forecasting models, they did not explore methods to identify underlying relationships between truck activity and commodity carried. Such a relationship is key in travel demand forecasting models that rely on predictions of commodity consumption and production trends using economic forecasts.

To better link the estimation of commodity production and consumption to freight flows, specifically truck volumes, it is key to measure truck flows by commodity carried. Uniquely, truck flows distinguished by commodity carried provide deep insights into trip patterns and can highlight potential sources of freight bottlenecks. For long-haul trips, average trip length (ATL) varies by commodity carried (*11*, *12*). However, ATL is the only trip characteristic available from most surveys, for example, the Vehicle Inventory and Use Survey (VIUS) used for freight analysis, and it is likely that other trip characteristics vary by commodity carried (*12*). Unfortunately, being a national inventory, VIUS does not cover daily trip patterns, trip chains, or shorter trips resulting from needs for rest breaks, fuel, and so forth. Therefore, it is necessary to identify key freight operational characteristics at smaller levels of geography (e.g., state, regional, or county) that can be used for more comprehensive freight planning.

To address the critical need for methods to extract operational characteristics from mobile sensors data, this study develops and combines three transferable heuristic algorithms to identify stop characteristics and trip characteristics from truck GPS data: (1) "stop identification" to aggregate pings (latitude, longitude, timestamp data points) into freight activity stops, that is, pick-up/drop-off or rest stops, (2) "path identification" to convert sparse pings into complete, fully connected paths on a dense transportation network, and (3) "trip identification" to extract operational characteristics by combining results of stop-identification and path-identification algorithms. The algorithms were applied to a sample of 338 million GPS pings (e.g., latitude, longitude, timestamp datapoints) collected from major trucking companies through on-board GPS units. The data represented 358,092 unique trucks during an 8-week sample period in Arkansas (i.e., a statewide region). The dataset contains a unique but anonymous truck ID so that location records for the same truck can be grouped. The data does not contain or describe key characteristics of the trucks such as vehicle type, commodity carried, and purpose of travel. Finally, to identify the operational characteristics that can be linked to commodity carried, a multinomial logit (MNL) model is applied. Application of these approaches to mobile sensor data enables such sources of big data to be used effectively for TPM while maintaining the anonymity of those sources.

Background

This section reviews prior research focused on heuristic approaches, methods, and models that were used to extract freight operational characteristics from large streams of truck GPS data.

Stop Identification

The premise of stop identification is to determine the locations of potential activity-related stops (e.g., fuel stops, rest stops, and pick-up/delivery for freight trucks) within large streams of GPS data points (also known as "pings"). Simple algorithms consider a stop to be the

location where the vehicle's instantaneous speed is recorded as zero. Minor modifications may assume the speed to be below a given threshold, say 3 mph. However, these simple approaches may overestimate the number of stops made by a vehicle by not grouping consecutive (redundant) pings representing zero or low speed into a single stop. In short, effective algorithms should combine consecutive, low-speed pings into clusters, determine the physical location of the stop within the cluster, and calculate a stop duration considering all pings in the cluster.

Existing stop-identification algorithms used geographic bounding boxes and rule-based approaches to define stop clusters (*5*, *13–17*). Greaves and Figliozzi developed a stop-identification algorithm for commercial vehicles and used the time difference between GPS-to-satellite communications to determine if the vehicle was stopped (*13*). The algorithm considered a time threshold of 4 min and a geographic distance threshold of around 20 ft (6 m) to identify a stop. If a vehicle repositioned by less than the defined threshold, regardless of the time elapsed, they performed a manual inspection to check whether it was a short stop. However, relying on manual inspections is not feasible for a large dataset. McCormack et al. identified delivery stops by defining a threshold of 3 min for dwell time (i.e., duration of a vehicle's engine as off or idle status) (*14*). To avoid redundant GPS pings of an idle truck, their algorithm removed data points where the distance between two consecutive pings was less than 65 ft (about 20 m). Though this filtered out false trips, it removed data that could be significant for deriving freight operational characteristics such as service times (i.e., the time for a truck to unload and start the next trip).

Holguín-Veras et al. used a mechanistic procedure to identify freight activity stops from GPS data (*15*). The driving pattern of freight trucks was the base of their procedure. After implementing the approach in three international case studies, they found that their approach can identify freight activity stops with an average accuracy of 98.6% (*15*). Alho et al. compared different algorithms used for stop-to-tour assignment and tour-type/chain identification (*16*). For their stop-to-tour assignment algorithm, they considered the "base" location of a trip as the start/end of a tour. For tour-type/chain identification algorithms, they considered the predominant tour-type identified for 1 day as well as the average number of stops per tour by stop type. After comparing their algorithms in an international case study, they found that the predictions of tours, tour types, and tour chain types were dependent on the assumptions made and the methods used for data processing (*16*).

The stop-identification algorithm developed by Camargo et al. expanded on the abovementioned research by using coverage and space-mean-speed (SMS) in addition to dwell time to define a stop (*5*). After grouping pings for which the travel speed between consecutive GPS records was less than 5 mph (8 km/h), they assessed the coverage of the set of pings. If a truck traveled less than 0.5 mi (about 800 m) between stops, pings were combined to represent a single stop. The geometric center of the stop cluster was defined as the stop location. The stop-identification method developed by Camargo et al. was used in this work with several modifications to ensure transferability among datasets, for example, metropolitan versus statewide spatial coverage (*5*).

Path Identification

Path identification, also known as map matching, refers to the process of identifying the network link that corresponds to each GPS ping (latitude, longitude, and timestamp data triples). Existing map-matching algorithms were developed based on the premise of assigning the pings to their closest network link and then connecting disparate links via shortest-path-finding algorithms (*5*, *18*, *19*). Giovannini's algorithm reconstructed routes from low-sample-rate GPS data, for example, around 1 mi between pings, using a Bayesian approach (*18*). Quddus and Washington developed a weight-based shortest path and vehicle trajectory aided map-matching algorithm to determine the network link corresponding to each GPS ping based on proximity, among other factors, for a sparse road network (*19*).

With temporally sparse GPS data, simple matching of the GPS ping to the closest link may not result in a complete and connected path. For example, many network links may be traversed between consecutive pings if the pings are recorded only every 15 min and a vehicle is traveling at highway speeds of 55 mph; there would be gaps when constructing the complete path of the truck from origin to destination. Camargo et al. addressed this gap by determining a fully connected complete path between sparse pings by applying shortest-path algorithms (*5*). The map-matching algorithm developed by Camargo et al. was used in this paper with several modifications to ensure route accuracy for a denser road network (*5*).

Freight Operational Characteristics from Mobile Sensor Data

Identifying stops and routes from GPS data allows us to compute network volumes and link/corridor speeds, identify bottlenecks, and estimate many other performance metrics for TPM. For freight-oriented TPM, it is also important to differentiate performance measures by operational characteristics such as trip type (e.g., long-

haul and short-haul trip), stop and trip purpose (e.g., rest, pick-up delivery, pass through), and industry served to enhance our understanding of economic impacts tied to freight movements.

Yang et al. characterized freight delivery stops from other types of stops using GPS data and a support vector machine (SVM) method (*20*). An SVM is a machine learning method commonly used as a pattern classifier. An SVM represents training data in a transformed feature space so that the points can be separated by a hyperplane with the largest margin separating a pair of classes. Test data are then mapped into the same space and predicted to belong to one side of the separating hyperplane. Three parameters, for example, stop duration, the distance to the center of the city, and the binary distance to a stop's closest bottleneck, served as input features of the SVM and produced minimal error of 0.2% (*20*). Based on trip length and number of trips derived from truck GPS data, Zanjani et al. distinguished light-duty local delivery trucks from long-haul operations using heuristic approaches (*6*). A local delivery truck was characterized as making more than five trips per day, none more than 100 mi in length. In combination with a driver survey, Jing analyzed stop purpose, stop duration, and stop time of day (*21*). Her study found four types of overnight, urban truck tours: one pick-up followed by one delivery, multiple consecutive pick-ups followed by one delivery, one pick-up followed by multiple consecutive deliveries, and multiple consecutive pick-ups followed by consecutive deliveries. Akter and Hernandez developed a supervised machine learning model to predict industry groups from anonymous GPS data (*22*). Their model allows large streams of truck movement data to be leveraged for freight travel demand forecasting.

None of the studies mentioned above were aimed at identifying or deriving freight operational characteristics that distinguish freight daily activity patterns by commodity carried or industry served. Knowledge of industry served can be used to estimate economic impacts associated with performance measurements, prioritize critical freight corridors according to key industries, and relate changes in economic conditions to transportation system performance. This paper relates operational characteristics defined from stop- and path-identification algorithms to trip type, stop and trip purpose, industry-associated trip chaining, or activity patterns.

Methodology

The methodology consists of four key approaches: (1) establishing consistency and relevancy of GPS data to improve algorithm performance, (2) modification of stop- and path-identification algorithms, (3) derivation of truck operational characteristics, and (4) development of an MNL model to distinguish trucks by industry served.

Data Consistency and Relevancy

Most commonly used truck GPS data sources require preprocessing to remove noise and other inconsistencies (*5*). Considering large-scale data, it is not possible to manually remove inconsistent records. Therefore, this paper presents an algorithmic data validation approach to improve data consistency and relevancy. The "consistency and relevancy" (CR) approach identifies a complete truck record for input into the stop-identification and path-identification algorithms. Complete truck records were defined as those that represented an over-the-road truck movement with logical start and end positions, speeds, and accelerations.

The CR algorithm identified the inconsistent truck trajectories and flagged those records for further analysis. First, the acceleration/deceleration rate of each truck for each pair of consecutive pings was calculated and pings that produced acceleration/deceleration rates above a predefined threshold of 2.24 ft/s^2, corresponding to 85th percentile average acceleration rate of heavy trucks, were removed (*23*). Next, the total number of pings corresponding to each truck record was calculated and truck records that had fewer pings than the threshold count (p_{count}) were removed. Then, the SMS and travel time between each consecutive pair of pings were calculated. Truck records were removed when the calculated SMS exceeded the speed limit (s_{max}) for a threshold time (t_{max}). Lastly, the geographic coverage area for each truck was calculated and any truck records that had a smaller geographic coverage area than the threshold area (c_{max}) were removed. Geographic coverage was defined as the diagonal of the rectangular bounding box that surrounds all pings of a truck (*5*). The algorithm flagged and removed 11% of the truck records so that the final dataset included about 300,000 unique truck records.

Stop- and Path-Identification Algorithms

The stop-identification algorithm developed in this paper was modified from Camargo et al. (*5*). Camargo et al. calibrated and validated their algorithm using truck GPS data from a metropolitan area of approximately 9,000 sq. mi (*5*). Comparatively, the study area of this paper encompasses the state of Arkansas, U.S., an area of approximately 53,000 sq. mi. In addition to the increase in geographic scale, there are complexities related to freight activity that require specific modifications to the original algorithm developed by Camargo et al. (*5*). Therefore, the values of the algorithm parameters were tailored to the Arkansas GPS data to identify stops more

Table 1. Parameter Values for the Stop-Identification Algorithm

Stop parameters	Original value	Modified value
Speed	5 mph (8 km/h)	3 mph (4.8 km/h)
Time	5 min (300 s)	5 min (300 s)
Coverage area	0.5 mi (0.8 km)	0.2 mi (0.3 km)

accurately. Table 1 juxtaposes the original and modified parameter values for the stop-identification algorithm. To arrive at the adjusted values, the following was done. Stop identification was applied to all sample periods (358,902 trucks) using the original parameter values. Then, truck records were sampled for manual verification of the identified "stops." Stops identified through the algorithm were verified by comparing with Google Earth satellite imagery that showed business names, parking facilities, and so forth, and adjustments were made to the parameters such that identified stops were at reasonable locations in relation to land use.

Stops were extracted from the set of valid truck records identified through the CR algorithm. The stop-identification algorithm calculated the SMS (s_j) between consecutive pings (p_{j-1} and p_j). If the SMS was less than a defined threshold speed (s_{min}) (i.e., 3 mph, as used in this paper) for at least a threshold time (t_{min}) (i.e., 5 min, as used in this paper), the algorithm continued by calculating the speed between the next pair of consecutive pings. Next, a series of the pings that passed the speed and time criteria, $\{p_j, p_{j+1}, \ldots, p_J \mid s_j \leqslant s_{min\ AND}\ t_j \geqslant t_{min}\}$ were collected. Next, the total stop coverage (c_T) and the total stop duration (t_{TQ}) for all consecutive pings from the series were calculated (Equations 1 and 3). If the total coverage for the series of the pings was less than c_{max} (i.e., 0.2 mi, as used in this paper), then the series was considered as a stop cluster (Q) (Equation 2). Although Camargo et al. specified the geographical center of the stop cluster (Q) as the stop location of the cluster, in this paper it was noted that the geographical center could be incorrect occasionally (e.g., in the middle of a road) (*5*). Therefore, the first identified stop's location (l_j) was used as the stop location for the stop cluster (Q). Ultimately, the algorithm produces a set of stop locations (i.e., pick-up/delivery stops, rest or fuel stops) along with stop time of day, stop duration, and stop coverage for each truck record.

c_T = geographical coverage of all consecutive stops (1)

$$Q = \{p_j, p_{j+1}, .., p_J \mid c_T < c_{max}\} \quad (2)$$

$$t_{TQ} = \sum_{j=1}^{J} t_j \quad ; \forall j \in Q \quad (3)$$

where

c_T = diagonal of the rectangular bounding box that surrounds all consecutive stops,

Q = a stop cluster of consecutive stop pings,

p_j = GPS pings, where $j = 1, \ldots, J$,

t_j = calculated travel time from current (p_j) and previous (p_{j-1}) timestamp, where $j = 1, \ldots, J$,

t_{TQ} = total stop duration for a series of consecutive stops, Q.

The path-identification algorithm identified the set of links that comprised the complete path between consecutive pings. Because of the temporal coarseness of the GPS pings and the density of the network links, this was a critical step in determining truck volumes along each link of the transportation network. For example, trucks traveling at 60 mph traverse many links between pings, especially when links can be as short as 0.1 mi. Thus, simply matching pings to nearby links does not produce a connected path. Instead, a path-identification algorithm was created to reconstruct the complete and connected series of links from the ping data. Because of the temporal coarseness of the GPS pings and the density of the network links, this was a critical step in determining, at the aggregate level, the volume of trucks along each link in the network and, at the disaggregate level, the accurate distance and travel time for each truck record.

First, a spatial buffer (b) was created around each network link (r_l) and a graph was created. Next, likely links related to each GPS ping (p_j) were selected using the buffer. The link buffer helped to account for small, inherent inaccuracies in the GPS ping positions. After identifying likely links, the algorithm computed the path between vertices in the graph by discounting the cost of traversing those links. The link cost (i.e., travel time was used in this study) calculation for using those routes is shown in Equations 4, 5, and 6. Thus, a complete but shortest path for each truck can be estimated (Figure 1). Among three alternatives shown as feasible paths in Figure 1, the path-identification algorithm will choose the path that has the lowest link cost (i.e., travel time). If link cost for the 1st, 2nd, and 3rd alternatives are 5, 10, and 7 min, respectively, the algorithm will choose the 1st alternative as the final path. Like the stop-identification algorithm, modifiable parameters (i.e., buffer distance and link cost) for the path-identification algorithm were modified from Camargo et al. (Table 2) (*5*). A free-flow travel time, calculated from the posted speed limit and road distance, was used as the link cost in this study. Other variables (e.g., time of day, volume, capacity) that may affect travel time and cause road congestion were not considered in this algorithm. The link cost calculation for using those routes is shown in Equations 4, 5, and 6. Thus, a complete but shortest path for each truck can be

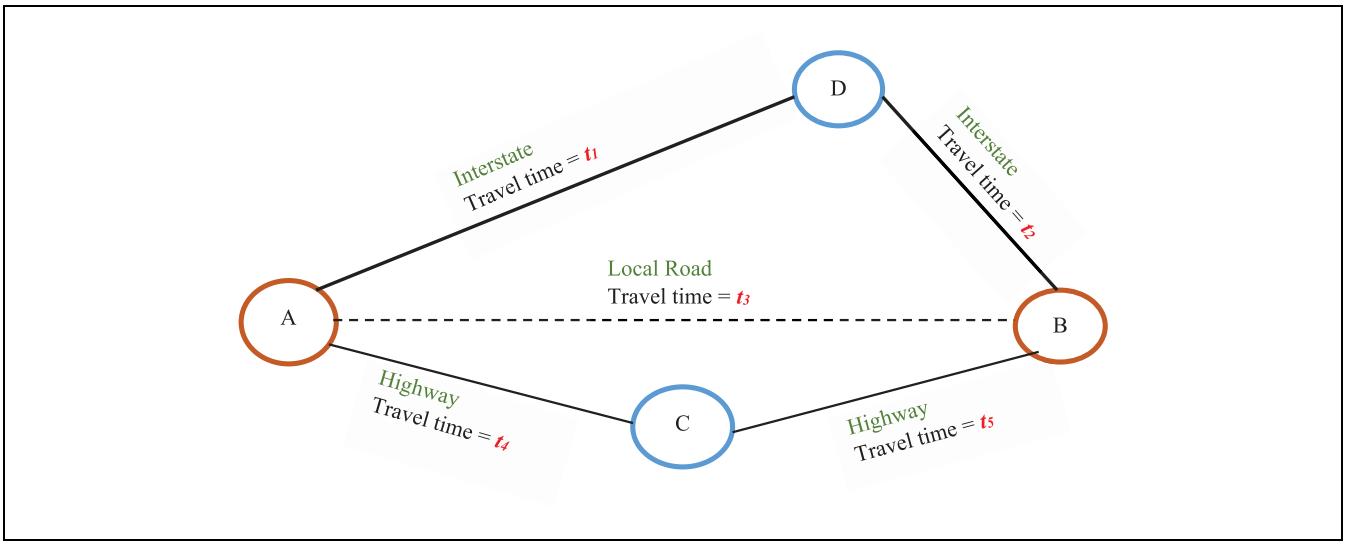

Figure 1. Shortest path considering travel times.

Table 2. Parameter Values for the Path-Identification Algorithm

Stop parameters	Original value	Modified value
Link buffer	1,654 ft	36 ft
Cost parameters	Link length	Travel time

Table 3. Path Identification Accuracy for Sample Period in Arkansas

Sample period	Accuracy (%)
February	88
May	87
August/September	87
November	86

estimated (Figure 1). The complete path can be a combination of highways and local roads.

We performed a sensitivity analysis for various threshold settings for speed, stop duration, and stop coverage by manually observing a set of truck records with diverse driving behaviors and determining at what threshold values most stops were reasonably captured by the algorithm. The threshold values are context-sensitive and, although they were informed by prior work, each parameter was tuned to the specific geography and land use characteristics of the study area (e.g., Arkansas). To validate the path-identification algorithm, a statistical verification procedure was developed and applied to a random sample of truck records. First, a buffer was created around the links found to be part of the complete path of a truck resulting from the path-identification algorithm. Next, the number of GPS pings for that truck contained within the link buffer was found. Then, the percentage of pings matched to a network link along the complete path was calculated. This value was referred to as the "path identification accuracy." This value represents the ability of the algorithm to capture the complete path of the truck. A path identification accuracy close to 100% is ideal. The path-identification algorithm with the modifications described in the previous sections has an average path identification accuracy of 87% for Arkansas (Table 3).

$$\text{1st Alternative: A} \rightarrow \text{D} \rightarrow \text{B} \qquad L_i = t_1 + t_2 \quad (4)$$

$$\text{2nd Alternative: A} \rightarrow \text{B} \qquad L_l = t_3 \quad (5)$$

$$\text{3rd Alternative: A} \rightarrow \text{C} \rightarrow \text{B} \qquad L_h = t_4 + t_5 \quad (6)$$

where

L_i = link cost for path 1 using interstates,
L_l = link cost for path 2 using local roads, and
L_h = link cost for path 3 using highways.

Derivation of Truck Operational Characteristics

The stop-identification algorithm identified sequential stops, and defined stops based on time, duration, and location. The path-identification algorithm reconstructed a path as a set of fully connected links defined by link identification number and timestamp. To derive operational characteristics, an algorithm was developed to merge results of the stop identification and path identification (see figures in the Appendix).

First, a serial number, s_j, was created for each stop of a truck based on the stop timestamp, t_j (i.e., time and date). Next, each pair of consecutive stops (s_j and s_{j+1}) were classified as a trip, m_j, that started with stop s_j and

Table 4. Example Results of Trip-Identification Algorithm

Trip ID	Stop pair	Stop time of day	Stop duration	Stop location	Road ID	Road length	Travel time	Travel speed	Road functional class
m_1	$\{s_1, s_2\}$	tod_{s1}	d_{s1}	l_{s1}	r_1	l_{r1}	t_{r2}	s_{r2}	Interstate
					r_2	l_{r2}	t_{r3}	s_{r3}	Interstate
					r_3	l_{r3}	t_{r4}	s_{r4}	Interstate
m_2	$\{s_2, s_3\}$	tod_{s2}	d_{s2}	l_{s2}	r_4	l_{r4}	t_{r5}	s_{r5}	Highway
m_3	$\{s_3, s_4\}$	tod_{s3}	d_{s3}	l_{s3}	r_5	l_{r5}	t_{r6}	s_{r6}	Highway
					r_6	l_{r6}	t_{r1}	s_{r1}	Local

ended with stop s_{j+1}. Thus, each trip was enveloped by two stops, that is, origin and destination. Stop information (i.e., stop time of day, stop duration, and stop location) of the origin stop were added to each trip.

However, some trips were not bounded by stops. This occurs when a portion of the trip or a stop is outside the boundary of the data sample. For example, for the sample used in this study, only pings within the Arkansas state boundary plus a 10 mi buffer (study area) were available. If a truck had a stop outside the study area, then it is not possible to observe that stop in the data sample. Likewise, it is not possible to observe the remainder of a trip outside the study area. These "open-ended" trips were still considered by bounding the trip by the state boundary, for example, the trip is defined from stop location to the state border and vice versa.

Secondly, path information (e.g., travel length, travel time, speed, and road link characteristics) was combined with stop information for each truck (example in Table 4). To combine path and stop data for each truck, the timestamp (t_k) associated with usage of road (r_k) was compared with the stop timestamps (t_j) for trip (m_j) such that t_k is greater than t_j and smaller than t_{j+1}. Later, trip length and trip duration are calculated from the combined table (Equations 7 and 8).

$$T_{m_j} = \sum_{k=1}^{n} t_{r_k} \tag{7}$$

$$L_{m_j} = \sum_{k=1}^{n} l_{r_k} \tag{8}$$

where

T_{m_j} = trip duration for trip m_j,
t_{r_k} = travel time for crossing a road link r_k,
n = number of road links in trip m_j,
L_{m_j} = trip length for trip m_j, and
l_{r_k} = length of road link r_k.

By merging the stop and path identification results, trip chains can be observed, and thus freight operational characteristics can be derived. Based on a review of the literature and the available data, 12 operational characteristics can be defined which can be aggregated into five groups (Table 5). First, stops are categorized based on stop duration into three categories: less than 30 min, 30 min to 8 h, and more than 8 h. These ranges coincide with hours of service (HOS) regulations for required rest breaks (*24*). For trip length and duration, trips are categorized based on general breakpoints found in the literature defining long and short haul trips. Time of day and total number of daily stops are also considered as important operational characteristics.

Table 5. Operational Characteristics by Group and Type

Feature group	Features	Variable type
Stop duration	1) Less than 30 min 2) 30 min to 8 h 3) More than 8 h	Discrete
Trip length	4) Less than 30 mi 5) 30 mi to 100 mi 6) More than 100 mi	Discrete
Trip duration	7) Less than 1 h 8) 1 h to 4 h 9) More than 4 h	Discrete
Time of day	10) Proportion of daytime stops (6 a.m. to 6 p.m.) to all stops 11) Proportion of night-time stops (midnight to 6 a.m. and 6 p.m. to midnight) to all stops	Continuous
Daily stop	12) Total number of stops in a day	Discrete

Development of a Multinomial Logit (MNL) Model

An MNL model was estimated to define associations between operational characteristics and the probability that a truck was transporting a certain commodity. The resulting model allows for the prediction of the commodity carried by a truck, as it is assumed that the observed stop and trip characteristics are, in part, the result of the

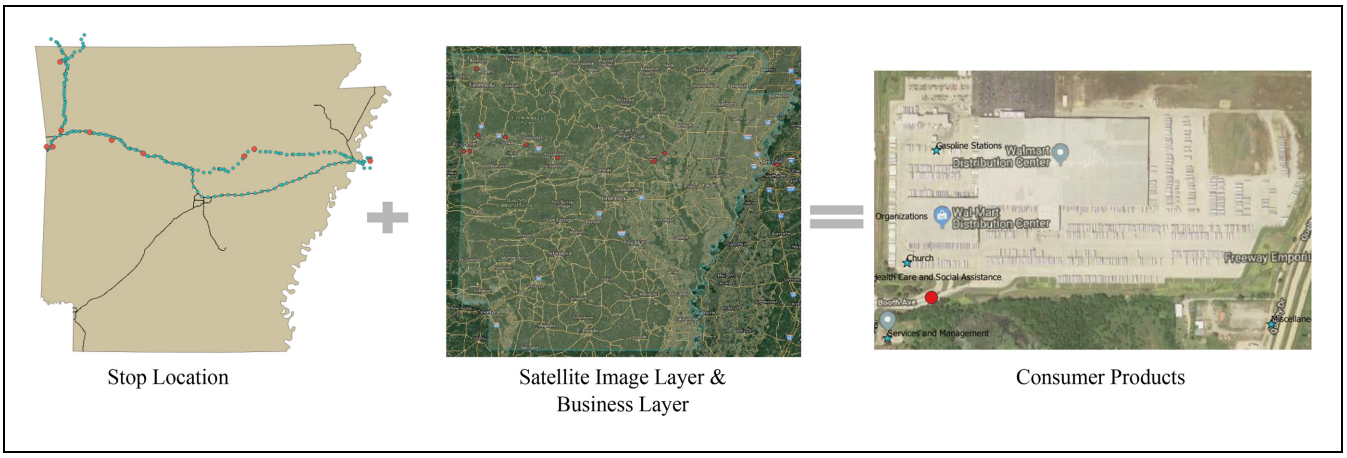

Figure 2. Prediction of carried commodity of a truck.

commodity being transported. Thus, the probability of a truck transporting commodity i can be calculated as:

$$P(i \backslash C_n) = P_r\left(U_{in} \geqslant U_{jn}\right), \forall j \in C_n \quad (9)$$

where

U = utility of the given alternative, and

C_n = {farm products, manufacturing, mining, chemicals, miscellaneous mixed, and pass-through}.

In this interpretation, the "utility" of alternative i can be calculated based on the stop and trip characteristics as:

$$U_{in} = \beta_{in} x_{in} + \varepsilon_{in} \quad (10)$$

where

U_{in} = estimated "utility" of alternative (commodity) i for driver/truck n,

x_{in} = observed stop and trip characteristics,

β_{in} = vector of coefficients of the variables, and

ε_{in} = random component, for example, unobserved or unmeasurable.

Under the assumption of the MNL model and based on the principle of utility maximization, the choice probability for alternative i can be calculated as:

$$P_n(i) = \frac{e^{Vin}}{\sum_{j \in Cn} e^{Vjn}}, \text{ for all } i \text{ in } j_n \quad (11)$$

where

$V_{in} = \beta_{in} x_{in}$

(All other terms previously defined.)

MNL Model Specification. A total of 11 of the 12 operational characteristics derived from the trip-identification algorithm were used (Table 5). To avoid multicollinearity, the "proportion of night-time stops" parameter was not included in the model.

Five commodity classes were considered in the model:

- manufactured goods
- farm products
- mining materials
- chemicals
- miscellaneous mixed

Additionally, "pass-through" trucks were considered as a "commodity." This was a necessary addition, as pass-through trucks represent unique operational behaviors that are not tied to specific commodities but are, nonetheless, included in the data sample because of only partial observation of the trip chain within the state boundary. The commodity category "farm products" was chosen as the base category.

MNL Model Estimation. Labeled data is needed to estimate the MNL model. In this application, labeled data refers to assigning a commodity carried to each observed truck trip contained in the GPS sample. To do this, a "ground truth" dataset was created using manual processes and comprised 2,064 truck trips. The assumption of commodity carried was based on a detailed examination of the truck trip and stops against aerial images depicting business and land uses, for example, Google Satellite images, along with North American Industry Classification System (NAICS) codes identified for the businesses where the stops were located (Figure 2). NAICS codes are adopted by the U.S. federal statistical agencies to classify business establishments. When manual observation of the business type could not fully distinguish the possible commodity carried, the NAICS code was referenced. For example, a business coded as NAICS 31-33 was considered "manufactured goods" while those classified as NAICS 339000 were identified as "miscellaneous mixed." Five commodity groups were clearly

Table 6. Change in Operational Characteristics Based on Commodity Groups

Features		Commodity groups					
Group	Description	Manufactured goods	Farm products	Mining materials	Chemicals	Misc. mixed	Pass-through
Stop duration							
Short break	Less than 30 min	0.01	−0.25***	−0.18***	−0.07	0.46***	0.03
Pick-up/delivery	30 min to 8 h	0.05	−0.02	0.18***	−0.40***	0.07	0.12
Long rest break	More than 8 h	0.47***	0.19*	0.52***	0.22	−0.51**	−0.88***
Trip length							
Short trip length	Less than 30 mi	0.38***	0.41***	0.27***	−1.37***	0.16	0.15
Medium trip length	30 mi to 100 mi	0.23***	0.48***	0.52***	−0.41**	−0.21	−0.60**
Long trip length	More than 100 mi	−0.08	−0.97***	−0.28***	1.53***	0.08	−0.28
Trip duration							
Short-trip duration	Less than 1 h	−0.44***	−0.02	−0.08	1.40***	−0.55***	−0.32
Medium-trip duration	1 h to 4 h	0.29***	0.57***	0.44***	−0.47***	−0.35***	−0.48**
Long-trip duration	More than 4 h	0.67***	−0.64**	0.15	−1.18***	0.93***	0.06
Time of day							
Daytime hours	6 a.m. to 6 p.m.	1.39***	0.28	−0.72***	−0.23	−0.16	−0.57
Daily stop							
Total stops	Total number of stops in a day	0.02***	−0.02***	0.00	−0.04***	0.01	0.02***
Log-likelihood: −1,993.31.							

Note: Misc. = miscellaneous.
*significant at 90% confidence level;
**significant at 95% confidence level;
***significant at 99% confidence level.

distinguishable in the data. To be clear, this is a required procedure for model development and not part of the proposed classification algorithm. Using this approach, the anonymity of the data is still maintained. Commodity groups were treated as the dependent variables and operational characteristics were treated as the independent variables of the MNL model.

Maximum likelihood estimation (MLE) was used to estimate the coefficients of the MNL model (*25*, *26*). At the 95% confidence level, stop duration, trip length, trip duration, stop time of day, and total number of daily stops were found to be significant parameters in predicting commodity carried (Table 6). The model data was split 75% for training and 25% for testing using the stratified random sampling with stratification based on commodity group. The overall accuracy of the model in relation to the correct classification rate (CCR) was 61% for the training data and 57% for the test data. The CCR is the ratio of the correctly classified responses to the total number of responses (Equation 12). This is also referred to as "recall" (Equation 12). CCR ranges from 79% for manufacturing to 0% for pass-through and is generally correlated with the volume of samples (Table 7). Since the model output could be used to predict the volume of each commodity group, we also present the volume accuracy of the classifications. Volume accuracy for each commodity group (i) is expressed as the difference in predicted and actual volume relative to the actual volume (Equation 13, all terms defined in Table 7).

$$CCR^i = 100\% \times \frac{P^i}{V_A^i} \tag{12}$$

$$\text{Volume Accuracy}^i = 100\% - \frac{\left|V_A^i - V_P^i\right|}{V_A^i} \tag{13}$$

Discussion

Knowing the commodity carried by a truck provides insight into its operational characteristics, for example, number of stops, trip length, and time of day travel patterns. Conversely, knowledge of operational characteristics can be used to understand the commodity carried by a truck. Because operational characteristics can be derived from GPS data, but commodity carried cannot be observed, the approach explained in this paper was to use heuristic methods to derive operational characteristics from GPS data and then relate those characteristics to commodity carried via an MNL model.

According to the estimation results, stop time of day, stop duration, trip length, and trip duration are found to be significant operational characteristics predictive of commodity carried. For instance, the model estimates that if the number of short break (less than 30 min)

Table 7. Cross-Classification Matrix and Model Statistics of Test Data

		Chemicals	Farm Products	Manufacturing	Mining	Misc. mixed	Pass-through	No. samples V_A^i	No. correct predictions P^i	Correct classification rate (CCR) (%)
Actual commodity classification	Chemicals	2*	0	5	1	0	0	8	2	25
	Farm products	0	56*	31	38	0	0	125	56	45
	Manufacturing	2	12	166*	31	0	0	211	166	79
	Mining	1	36	53	66*	0	0	156	66	42
	Misc. mixed	0	0	7	0	2*	0	9	2	22
	Pass-through	0	1	4	2	0	0*	7	0	0
Total volume predicted, V_P^i		5	105	266	138	2	0	516	292	57
Volume accuracy (%)		63	84	74	88	22	0	na	na	na

Note: Misc. = miscellaneous; na = not applicable.
*Gray-Shaded Cells = Correctly classified commodity
Gray-Shaded Cells = Total

increases by one, the log-odds of carrying miscellaneous mixed goods will increase by 0.46 and the log-odds of carrying farm products will decrease by 0.25. It also indicates that the probability of carrying miscellaneous mixed commodity will increase by 55% while the probability of carrying farm products will decrease by 24%. This denotes that trucks carrying miscellaneous mixed products have higher number of short breaks compared with trucks carrying farm products, which follows intuition that trucks carrying miscellaneous mixed goods are making stops at many different establishments in a less-than-truckload fashion. The model also estimates that, if the number of short trips (less than 30 mi) increases by one, the log-odds of carrying farm products will increase by 0.41, while the log-odds of carrying chemicals will decrease by 1.37. In other words, the probability of that truck carrying farm products will increase by 31% while the probability of it carrying chemicals will decrease by 78%. It indicates that trucks carrying farm products have a higher number of short-length trips compared with trucks carrying chemicals. This can be intuitively supported by noting that farm products include transporting animal feed from local producers to farms, as well as live animals to local processing plants. Additionally, the model finds that the stop time of day factor is positively significant for manufactured products and negatively significant for mining materials. It indicates that trucks carrying manufactured products have a higher number of stops during daytime (6 a.m. to 6 p.m.) compared with trucks carrying mining materials. Again, this reflects intuition, in that manufactured product deliveries must follow daily operating hours of factories and stores.

The data used in this work represents a sample of trucks covering a statewide region. On average, this sample represents approximately 10% to 15% of the total population of trucks. From that sample, we manually identified the industry of 1,584 trucks to train our model (75%) and 516 trucks to test the model (25%). This is a common split of training and test data for machine learning applications. Further, we use stratified random sampling with stratification based on commodity group to ensure the training and testing data are representative of the larger population of trucks. It can be argued that more training data will result in a stronger model while more testing data can better show how performance varies by class. However, the approach taken in this paper to gather labeled samples required a manual, time-consuming process and thus was restricted to 2,064 trucks. There is no known data set in the public space that links commodity carried to GPS record and, thus, it is a challenge to gather labeled training instances for model development. For future expansion of this work, researchers can partner with trucking companies to potentially gather paired commodity and truck movement data for model-training purposes.

Conclusion

Although big data such as that from GPS is increasingly plentiful, without efficient heuristic methods to extract relevant performance measures it is not possible to fully leverage this valuable data source. Methods to derive stop duration, trip length, trip duration, and stop time of day allow us to identify freight activity patterns from big data sources and to link those patterns to commodity carried. While deriving operational characteristics from big data allows us to develop more ubiquitous transportation performance metrics, the link between operational characteristics and commodity carried serves as critical input for freight demand forecasting that hinges on economic forecasts of commodity production and consumption (*11*).

The methodology presented in this paper consists of spatial heuristics to identify stop clusters and complete paths of individual trucks from timestamped latitude-longitude points gathered from GPS devices on board trucks. After deriving stop and path, trip chains—for example, sequences of stops and trips—can be observed. Statistical approaches, namely MNL models, were employed to determine how operational characteristics such as stop time of day and duration, relate to commodity carried. The MNL model identified that stop duration, number of total daily stops, stop time of day, trip length, and trip duration were significant characteristics that could be used to predict commodity carried.

Although the CCR of the predictions is 57%, and several of the commodity class predictions are above 40%, the pseudo-R-squared of the estimated MNL model of 29%, a general description of the goodness of fit, indicates that there is a room for improvement. This can be attributed to several factors. First, MNL estimation assumes a linear-in-parameters specification such that operational characteristics should be linearly related to commodity carried. This assumption may not hold true. Advanced machine learning methods such as *K*-means clustering, random forest, and SVM models can better identify patterns, especially non-linear patterns, from large and noisy data like GPS pings (*27*). Therefore, machine learning models are likely more appropriate for this application but would require more ground truth (training data). Second, MNL specification requires a complete choice set to be specified. This paper considered only five commodity groups plus a sixth group representing pass-through movements. This is not a complete choice set but was limited by data ground truth procedures necessary to maintain confidence in the labeled data. Future work should expand the set of commodities which should also improve cross-classification errors.

The results of this paper can guide public sector engineers and planners to achieve the TPM goal-setting initiatives and requirements set forth in federal transportation legislation.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: T. Akter, S. Hernandez; data collection: T. Akter; analysis and interpretation of results: T. Akter, S. Hernandez, P. Camargo; draft manuscript preparation: T. Akter. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors thank the Arkansas Department of Transportation (ARDOT) for sponsoring the project that led to this paper. ARDOT Grant Reference No.: TRC 1702.

ORCID iDs

Taslima Akter https://orcid.org/0000-0001-9585-7346
Sarah Hernandez https://orcid.org/0000-0002-4243-1461
Pedro V. Camargo https://orcid.org/0000-0001-9613-2777

Supplemental Material

Supplemental material for this article is available online.

References

1. Roorda, M. J., R. Cavalcante, S. McCabe, and H. Kwan. A Conceptual Framework for Agent-Based Modelling of Logistics Services. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 46, No. 1, 2010, pp. 18–31. https://doi.org/10.1016/j.tre.2009.06.002.
2. CPCS. NCFRP 49 [Final]: Understanding and Using New Data Sources to Address Urban and Metropolitan Freight Challenges. 2018. http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID = 3593.
3. Bassok, A., E. D. McCormack, M. L. Outwater, and C. Ta. Use of Truck GPS Data for Freight Forecasting. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
4. Kuppam, A., J. Lemp, D. Beagan, V. Livshits, L. Vallabhaneni, and S. Nippani. Development of a Tour-Based Truck Travel Demand Model Using Truck GPS Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
5. Camargo, P., S. Hong, and V. Livshits. Expanding the Uses of Truck GPS Data in Freight Modeling and Planning Activities. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2646: 68–76.
6. Zanjani, A. B., A. R. Pinjari, M. Kamali, A. Thakur, J. Short, V. Mysore, and S. F. Tabatabaee. Estimation of Statewide Origin–Destination Truck Flows from Large Streams

of GPS Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2015. 2494: 87–96.
7. Sharman, B. W., and M. J. Roorda. Analysis of Freight Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2011. 2246: 83–91.
8. Liao, C. *Using Archived Truck GPS Data for Freight Performance Analysis on I-94/I-90 from the Twin Cities to Chicago*. University of Minnesota Center for Transportation Studies, 2009. https://conservancy.umn.edu/handle/11299/97668.
9. Ma, X., E. D. McCormack, and Y. Wang. Processing Commercial Global Positioning System Data to Develop a Web-Based Truck Performance Measures Program. *Transportation Research Record: Journal of the Transportation Research Board*, 2011. 2246: 92–100.
10. Zhao, W., E. McCormack, D. J. Dailey, and E. Scharnhorst. Using Truck Probe GPS Data to Identify and Rank Roadway Bottlenecks. *Journal of Transportation Engineering*, Vol. 139, No. 1, 2013, pp. 1–7. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000444.
11. Beagan, D., D. Tempesta, and K. Proussaloglou. *Quick Response Freight Methods*. 2019. https://ops.fhwa.dot.gov/publications/fhwahop19057/fhwahop19057.pdf.
12. Evans, D. L., T. W. Kassinger, K. B. Cooper, and C. L. Kincannon. 2002 Vehicle Inventory and Use Survey. 2004. https://www.census.gov/library/publications/2002/econ/census/vehicle-inventory-and-use-survey.html.
13. Greaves, S. P., and M. A. Figliozzi. Collecting Commercial Vehicle Tour Data with Passive Global Positioning System Technology. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. 2049: 158–166.
14. McCormack, E. D., X. Ma, C. Klocow, A. Currarei, and D. Wright. Developing a GPS-Based Truck Freight Performance Measures Platform. 2010. https://www.wsdot.wa.gov/research/reports/fullreports/748.1.pdf.
15. Holguín-Veras, J., T. Encarnación, S. Pérez-Guzmán, and X. Yang. Mechanistic Identification of Freight Activity Stops from Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2020. 2674: 235–246.
16. Alho, A. R., T. Sakai, M. H. Chua, K. Jeong, P. Jing, and M. Ben-Akiva. Exploring Algorithms for Revealing Freight Vehicle Tours, Tour-Types, and Tour-Chain-Types from GPS Vehicle Traces and Stop Activity Data. *Journal of Big Data Analytics in Transportation*, Vol. 1, No. 2–3, 2019, pp. 175–190.
17. Thakur, A., A. R. Pinjari, A. B. Zanjani, J. Short, V. Mysore, and S. F. Tabatabaee. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board*, 2015. 2529: 66–73.
18. Giovannini, L. A Novel Map-Matching Procedure for Low-Sampling GPS Data with Applications to Traffic Flow Analysis. 2011. https://doi.org/10.6092/unibo/amsdottorato/3898; https://www.openaire.eu/search/publication?articleId=od_____1754::2e76bee797112fda11280f4851def321.
19. Quddus, M., and S. Washington. Shortest Path and Vehicle Trajectory Aided Map-Matching for Low Frequency GPS Data. *Transportation Research Part C: Emerging Technologies*, Vol. 55, 2015, pp. 328–339. https://doi.org/10.1016/j.trc.2015.02.017.
20. Yang, X., Z. Sun, X. J. Ban, and J. Holguín-Veras. Urban Freight Delivery Stop Identification with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2014. 2411: 55–61.
21. Jing, P. *Identifying and Modeling Urban Truck Daily Tour-Chaining Patterns*. Doctoral dissertation. Massachusetts Institute of Technology, Cambridge, 2018.
22. Akter, T., and S. Hernandez. Truck Industry Classification from Anonymous Mobile Sensor Data Using Machine Learning. *International Journal of Transportation Science and Technology*, Vol. 11, No. 3, 2022, pp. 522–535.
23. Pline, J. L. *Traffic Engineering Handbook*, 5th ed. Institute of Transportation Engineers, Washington, D.C., 1999. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=009998668&sequence=000001&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.
24. FMCSA. Summary of Hours of Service Regulations. 2017. https://www.fmcsa.dot.gov/regulations/hours-service/summary-hours-service-regulations.
25. Ben-Akiva, M. E., S. R. Lerman, and S. R. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. Vol. 9. MIT Press, 1985.
26. Bunch, D. S. Maximum Likelihood Estimation of Probabilistic Choice Models. *SIAM Journal on Scientific and Statistical Computing*, Vol. 8, No. 1, 1987, pp. 56–70. https://doi.org/10.1137/0908006.
27. Caruana, R., and A. Niculescu-Mizil. An Empirical Comparison of Supervised Learning Algorithms. *Proc., 23rd International Conference on Machine learning, ICML '06*, Pittsburgh, PA, Association for Computing Machinery, New York, NY, 2006, pp. 161–168. https://doi.org/10.1145/1143844.1143865; http://dl.acm.org/citation.cfm?id=1143865.