FORECASTING OF ORIGIN-TO-DESTINATION REQUESTS FOR TAXIS USING DNN ALGORITHM WITH NYU DATABASE

R.B.SAROORAJ¹, S. PRAYLA SHYRY²

¹ Research Scholar, Sathyabama Institute of Science and Technology, Chennai, India
² Professor, Sathyabama Institute of Science and Technology, Chennai, India

Email: ¹ saroorajrb@gmail.com

ABSTRACT

Prediction of taxi requests has lately drawn increased much attention in research owing to its high applicability in massive intelligent transport systems. Most previous methods, on the other hand, focused solely on predicting taxi demand in origin areas, ignoring the analysis of target passengers' special circumstances. We believe it is inefficient to assign cabs to all areas in advance solely based on taxi origin request. This work studies a critical and fascinating task known as taxi origin-destination demand prediction, whose purpose is to forecast future cab requests across pairs in all areas. Determining the process to collect contextual data effectively in order to learn demand patterns. A novel methodology known as the Deep Neural Network (DNN) with Deep learning-based models is focused in this paper, which outperforms traditional machine learning methods in a variety of classification tasks, including origin and destination views. Extensive tests and analyses on a broad dataset clearly show that our DNN outperforms several other methodologies from literature.

Keywords: Taxi Origin-Destination, Spatial-Temporal Modeling, Deep Neural Network, DNN Algorithm, NYU Database.

1. INTRODUCTION

Being one among most prevalent modes of transportation for city dwellers, taxis have become deeply embedded in people's everyday lives. The ease of daily travel brought about by the taxis recently, have led to rapid growth in online platforms such as Uber² and Grab³. However, some wasteful activities still exist in this massive industry [1]. The key issue stems due to supply and demand disparity, created by incorrect taxi request predictions, resulting in large number of cabs assembling in crowded regions. This created an over supply meanwhile taxi distribution was extremely sparse in other remote areas. The solution to this problem is taxi demand prediction, estimating taxi requests in the future and assists in the allocation of taxis to each area ahead of time.

Estimating the cab requests has drawn a lot of attention as a critical role in intelligent transportation systems (ITS) and has had some notable successes [2]. Most current approaches, on the other hand, only model taxi demand at the departure point and forecast taxi demands in all areas or particular locations, overlooking the impact of the passenger’s destination. It is assumed that the passenger destination information is important for taxi pre-allocation systems. Without taking into account the distribution of passenger destinations, taxi pre-allocation systems dispatch taxis ahead of time solely based on expected taxi origin demand, which can result in the ensuing problems [3]:

Figure 1: Visualization map of taxi demands
Some drivers are only permitted to drive in certain areas due to regulations regarding management of the city, (like Beijing's policy of restricting driving). While being assigned to a certain area where the majority of passengers must travel to a location where drivers is limited, the driver may be unable to accept orders, potentially wasting resources.

Few drivers tend to transport commuters in areas where they are familiar. Few are unable to accept low-paying short-distance drives. The driver can refuse requests being if majority of their destinations in the driver's pre-assigned areas are outside to the area where they operate, or very near to some pickup locations.

In being dispatched to an area where the majority of commuters would be travelling to an unfamiliar region, the driver may take longer to drop the commuters at their destinations, even if GPS navigation is used. This would reduce the productivity of the taxi industry as well as customer satisfaction.

Taxi riders, on the other hand, can be found almost everywhere, and the forecasting systems for reading traffic flow designed might not be appropriate for pre-allocation of cabs all over the city. As a result, predicting the request of cabs in between two areas and optimizing the taxi allocation process becomes desirable [4].

A sophisticated technique that forecasts the action required for finding the origin and destination of a taxi has been constructed here. An effective taxi pre-allocation strategy is defined here to provide satisfactory demands to commuters and concurrently prevent all the issues upon finding familiarity with the pickup and destination points. The proposed task's main challenges are how to collect a variety of spatial-temporal contextual information in order to identify patterns pertaining to demand. In our research, we found that some spatially adjacent regions have request patterns that are identical (for example, the amount of taxi requests and trends in demand). Furthermore, even if two areas are separated geographically, demand patterns may be significant when they share similarities (both of them are residential districts). Cab request is a time-varying phenomenon driven by a variety of factors, including its present state and meteorology which modifies everytime, which is referred to as the temporal evolution context.

The rest of this article is structured as shown: Section II reviews some literature and interprets the taxi origin-destination request issue. It summarizes the Preliminaries of Taxi Demands in Section III. The proposed novel DNN algorithm with process methodology is defined in section IV. Conduct detailed experiments for findings and discussion in Section V. Finally, in Section VI, this proposed approach is put to rest.

2. LITERATURE SURVEY

This paper uses a novel DNN to estimate taxi requests in several areas, which thoroughly relates to estimation in taxi requests. In addition, the related works of these two categories of research will be extensively reviewed in the subsections that follow.

2.1. Prediction demand taxi

Because of its vast applicability in ITS, forecasting cab requests have indeed been studied in an extensive manner [5]. It proposed combining GPS signals onto histogram periodicity and forecasting demand using a Poisson Model and an ARMA model. [6] demonstrated a recommender that uses historical GPS to provide taxi drivers with precise positions for rapidly picking passengers up.[7] uses an improved ARIMA-based prediction model to forecast the spatiotemporal variations of passengers at a given hotspot. The taxi trajectories are needed for all of the methods above. The concept of trajectory-free prediction has recently gained popularity. In [8] proposed a groundbreaking study where a taxi-calling data as of several online taxi request sites are used to forecast taxi requests using a single linear regression model.

Deep learning techniques used for several computer vision functionalities [9] has recently prompted researchers to use a deep neural network to tackle this issue. [10] developed a neural network architecture for predicting the difference between taxi supply and demand using background data from multiple sources. [11] has suggested a model for forecasting future taxi requests in each area by employing sequence learning. The temporal and spatial connections have been accurately modeled [12] using a Deep Multi-View Spatial-Temporal Network architecture. In [13] used two hybrid deep learning architectures to estimate cab requests in event areas using time series and textual data.
Recently, [14] created an attention-based neural network to identify passenger transport requests by region, but it wasn't useful for predicting cab requests in between the region pairs.

### 2.2. Estimation for origin-destination

Identification of Origin and Destination points [15] assesses the path amongst the various endpoints of the investigated traffic network using the flow count and certain other findings among many traffic links. Static estimation and dynamic estimation are two types of research on this subject that have already been done. Static methods [16] predict average demand and treat traffic flow as time-independent, making them suitable for a long-term design and planning of transportations. On the other hand, dynamic methods [17] predict the time-variant path among each arrival and departure which can be employed for short-term route guidance and dynamic traffic assignment. The flow count of specific connections, as determined by appropriate positions (e.g., the joint in a main street, express lane, subway, toll booth in highway, and bus stations) and some prior knowledge are usually used in these works (i.e., the proportion among several origin-destination pairs). Upon consideration of many positions, the dimension of OD matrix becomes large and difficult to compute. Dimension reduction technology have been previously studied for solving this issue [18].

All of the methods above can only predict cab requests at every time in a given area or even at a particular location. Our tool, on the other hand, aims to forecast cab requests at each region, which can aid systems that could pre-allocate cabs in more efficiently allocating taxis. Furthermore, the proposed approach of novel DNN in this paper specifically captures more reliable taxi demand trends.

### 3. Preliminaries

We first define some notations in this section, and then use these notations to formulate the issue of identifying the source and destination requests of a cab.

#### 3.1. Partition of region

Rather than focusing on specific positions, we concentrate on predicting the source and destination requests of a cab between various regions in this study. Examples of means in which a city can be segmented into many regions in terms of changing granularities and semantic meanings are the zip code tabular and road network [19]. As in previous works, it divides a city across HW non-overlapping grid maps subject to latitude and longitude [20]. A different geographical area is represented by each rectangular grid in the city. The region on the grid map's ith row and jth column will be referred to as R in the following parts (i, j). This method could transform the actual taxi demand details directly into tensor or matrix.

### 3.2. Prediction of taxi demand for origin-destination

The aim of our work is to identify the issue related to finding the source and destination points of cab requests is to identify origin to destination requests for cab details Xt in t time interval, provided data up to time interval t−1. Since weather conditions have a significant impact on taxi demand, we also integrate meteorological historical information for this mission, and we refer to meteorological data in time intervals I as Mi. Section V covers the compilation and preprocessing of cab request information as well as meteorological data. As a result, our ultimate aim is to forecast Xt using historical request data (Xi|i = t n +1,...,t) and meteorological data (Mij|t = t n+1,...,t), where n being the time interval series duration [21].

### 4. Proposed Methodology

This section reports on the proposed DNN methodology for identifying issues of source and destination points regarding cab requests. As shown in Figure 2, novel DNN to understand the local spatial context of taxi demand from origin view and destination view.
a) **Data collection**
Data collection is the systematic method of collecting and evaluating information on variables of interest in order to address test hypotheses and determine outcomes. From April to September 2014, around six CSV files were identified with information containing pickup details of Uber in New York City. Every file has information for a specific month and contains the ensuing information: Time/Date: The pickup time and date for Uber, Lat: The Uber pickup's latitude, Lon: The Uber pickup location's longitude. TLC base company code [22], which is associated with the Uber pickup.

b) **Pre-processing**
The collected data is pre-processed, whose process entails converting raw information into a format that can be understood. Real-world data is often incomplete, unreliable, and/or deficient in specific habits or patterns, as well as containing numerous errors. Data reduction, transformation, integration, cleaning other preprocessing activities are among the most significant. The missing values are listed and mean values are substituted. The cleaned dataset is fed into the recurrent neural network model, which is trained using historical data such as date, time, pickup location, weather, and potential taxi demand is predicted [23].

c) **DNN**
An ANN comprising of various hidden layers in between the output and input layers is known as a Deep Neural Network (DNN). Compositional models are generated by DNN architectures, in which the object is represented as primitives in layered composition.
destination majority of DNN classifiers are structures that contain decisions based on the assumption that the amount of data in each category differs only slightly. A generalization effect arises when the impact of DNN exceeds a certain threshold, under which learning is manipulated towards that classification group [16]. The reliability of non-homogenous types of case data can be called into question [17]. In light of the above, it is essential to create a prediction method for DNN. The difficulty in removing features from a DNN is linked to the classifier's predictive efficiency. As previously mentioned, the output challenge induced by knowledge of various AI algorithms is that behavioral intelligence data is used concurrently with genetic variables, resulting in a problem [24-25].

\[ Rawdata \cap f_{DNN} \cap Resultdata \]

\[ Raw_v \cap A_k, A_{t_k},...,A_t \]
\[ Raw_v \cap B_k, B_{t_k},...,B_t \]
\[ Raw_v \cap A_k, A_{t_k},...,A_t, B_{t_k}, B_{t_{k+1}},...,B_t \]
\[ Raw_v \cap A_k, A_{t_k},...,A_t, D_{t_k}, D_{t_{k+1}},...,D_t \]
\[ Raw_v \cap A_k, A_{t_k},...,A_t, C_{t_k}, C_{t_{k+1}},...,C_t \]
\[ Raw_v \cap A_k, A_{t_k},...,A_t, k_{t_k}, k_{t_{k+1}},...,k_t \]
\[ C_k \cap C_{t_k} \cap k_{t_k}, C_{k_{t+1}},...,C_t \]
\[ C_k \cap C_{t_k} \cap k_{t_k}, C_{k_{t+1}},...,C_t \]
\[ Resultdata \cap A_1, A_t, 2, ..., A_t \]

where \( f_{DNN} \) is fitting function of the DNN structure. \( A \) denoted as the speed data, \( B \) denoted as steering data, \( C \) denoted as location data, and \( D \) denoted as driving data information. \( C_g \) denoted as longitude data, \( C_a \) denoted as latitude data, \( E \) denoted as altitude data. \( t \) denoted as present time. \( k \) denoted as historichorizon the data used to predict. \( m \) denoted as aspredict horizon.

d) Prediction
The networks have been able to learn more complex problems thanks to predictions. The term prediction refers to a norm that helps nodes make more general predictions, and bias nodes are used in almost all models. The network's loss function explains how far the network stands from providing superlative data estimations. The loss function's value just needs to be related that reduces as the predictions improve. The most straightforward method for prediction is to take the sum or mean of the divergences between the forecast and actual outputs, which will yield a number indicating how far off the predictions. To determine how well a network performs, it must have the ability to predict its outcomes during training [26].

e) Validation
In order to obtain reliable data, validation is added to the input testing. Validation appears to be an excellent option for checking network parameters. The issue in this report is that it contains a temporal aspect to it, making the three subsets split useful. Since the network should forecast future trips, it would be important to break the dataset into three chronological sections to see how well identifying values in the future works. The dataset must be large enough to not easily predict the test data by chance [27].

5. RESULTS AND DISCUSSION

This section establishes a benchmark in a large-scale for predicting source-destination requests for taxis. The task's evaluation metrics are then presented, and our proposed solution is compared to other advanced approaches. A detailed component review is performed finally, to establish the usefulness of every component in the proposed model:

5.1. NYC TOD datasets
There seem to be no publicly available datasets to forecasting origin-to-destination requests for cabs in areas that we are aware of. We further create a first standard for this mission, known as NYC-TOD, which will
be used to assess the outputs of all compared methods and encourage more applicable research. It is made up of two types: data regarding the forecasting of origin-to-destination requests for cabs and New York City meteorological data. The data from the previous sixty days is used as the testing collection, while all data prior to that is used as the training set. New York City (NYC) is one among the most affluent cities in the world, with a well-developed taxi industry. Since the majority of NYC taxi trips originate and end near the Manhattan borough, we have chosen Manhattan as the focus of this research. We divide Manhattan into a 155-grid map subject to longitude and latitude, as discussed in Section III. Each grid represents a geographical area measuring approximately 0.75km x 0.75km [28].

5.2. Evaluation parameter
We use the Rooted Mean Square Error (RMSE) and Mean Average Percentage Error (MAPE) as metrics to analyze the output of all methods, as described in previous works [29]:

\[
MAPE = \frac{1}{z} \sum_{t=1}^{z} \frac{|X_t - \hat{X}_t|}{X_t},
\]

\[
RMSE = \sqrt{\frac{1}{z} \sum_{t=1}^{z} |\hat{X}_t - X_t|^2}.
\]

Here z becomes the number of research samples, the expected taxi requests in the time interval t is given as Xt, and \(\hat{X}_t\) is the corresponding ground truth in time interval t, respectively [30]. The former task's MAPE and RMSE are referred to as OD-MAPE and OD-RMSE, while the latter task’s MAPE and RMSE are referred to as O-MAPE and O-RMSE.

<table>
<thead>
<tr>
<th>Method</th>
<th>OD-MAPE</th>
<th>OD-RMSE</th>
<th>O-MAPE</th>
<th>O-RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM (hidden Markov model) [30]</td>
<td>37.71%</td>
<td>1.94</td>
<td>45.01%</td>
<td>50.44%</td>
</tr>
<tr>
<td>CRF (conditional random field) [31]</td>
<td>35.46%</td>
<td>1.82</td>
<td>40.61%</td>
<td>40.97%</td>
</tr>
<tr>
<td>DT (decision tree) [32]</td>
<td>33.85%</td>
<td>1.65</td>
<td>33.79%</td>
<td>31.10%</td>
</tr>
<tr>
<td>RNN (recurrent neural network) [33]</td>
<td>33.86%</td>
<td>1.48</td>
<td>24.12%</td>
<td>25.47%</td>
</tr>
<tr>
<td>CNN (convolution neural network) [34]</td>
<td>27.04%</td>
<td>1.36</td>
<td>19.46%</td>
<td>21.55%</td>
</tr>
<tr>
<td>Proposed Novel DNN</td>
<td>26.30%</td>
<td>1.02</td>
<td>17.50%</td>
<td>18.99%</td>
</tr>
</tbody>
</table>

Table 1: Performance of Various Methodologies Using NYC-TOD Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Weekday</th>
<th>Weekends</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM (hidden Markov model) [30]</td>
<td>43.12%</td>
<td>40.86%</td>
</tr>
<tr>
<td>CRF (conditional random field) [31]</td>
<td>36.09%</td>
<td>35.17%</td>
</tr>
<tr>
<td>DT (decision tree) [32]</td>
<td>33.82%</td>
<td>31.58%</td>
</tr>
<tr>
<td>RNN (recurrent neural network) [33]</td>
<td>31.40%</td>
<td>30.18%</td>
</tr>
<tr>
<td>CNN (convolution neural network) [34]</td>
<td>28.99%</td>
<td>28.41%</td>
</tr>
<tr>
<td>Proposed Novel DNN</td>
<td>25.99%</td>
<td>27.50%</td>
</tr>
</tbody>
</table>

Table 2: The Comparison Of Weekdays And Weekends Of Source-Destination Cab Demand Forecasting By MAPE
requests in all areas by possessing knowledge of cab request patterns from historical data. Our model is trained and evaluated using records from 132 million taxi trips in New York City. The proposed model outperforms several advanced approaches on the challenge of taxi origin-destination request forecasting, with an OD-MAPE of 24.93 percent and an O-MAPE of 12.92 percent. It also allows us to expand the proposed DNN to forecast long-term taxi demands, and performs admirably. Meanwhile, more details can be added line periodic taxi requests and Points of Interest (POI) among each area. In doing so, the efficiency can be enhanced even more.

7. DECLARATION:

Ethics Approval and Consent to Participate:
No participation of humans takes place in this implementation process

Human and Animal Rights:
No violation of Human and Animal Rights is involved.

Funding:
No funding is involved in this work.

Conflict of Interest:
Conflict of Interest is not applicable in this work.

Authorship contributions:
There is no authorship contribution

Acknowledgment:
There is no acknowledgment involved in this work.

REFERENCES


