Predicting Taxi and Uber Demand in Cities Approaching the Limit of Predictability

Repalle Madhusudha

Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *Traditional taxi systems in cities often suffer from inefficiencies due to uncoordinated actions as customer demand changes. To predict the number of taxis that will emerge within the next time interval. we treat the taxi and Uber demand in each location as a time series, and reduce the taxi and Uber demand prediction problem to a time series prediction problem. we treat the taxi and Uber demand in each location as a time series, and reduce the taxi and Uber demand prediction problem to a time series prediction problem. We answer two key questions in this area. First, time series have different temporal regularity. Moreover, it reduces the level of passenger satisfaction because of the long wait times. The ability to predict taxi and Uber demand can help address the taxi-service inefficiency problem. With the deployment of the networked sensors and the widely used mobile phones in taxis, large amounts of information including location, time, number of passengers, weather, traffic, etc. can be collected in real time. This information provides opportunities to build an intelligent transportation system that is able to control and coordinate taxis at large scale and bring benefits to both taxi drivers and companies: taxi drivers can drive to high taxi demand areas, and ride-sharing companies (e.g., Uber) may re-allocate their vehicles using the surge pricing in advance to meet the passenger demand.*

Keywords: sharing economy; deep learning; predictive algorithm; predictability of time-series; data mining.

I. INTRODUCTION

Traditional taxi systems in cities often suffer from inefficiencies due to uncoordinated actions as customer demand changes [1]. In some areas passengers experience long waits for a taxi, while in others, many taxis roam without riders. This leads to profit loss for taxi drivers since vehicles are vacant when there is demand. Moreover, it reduces the level of passenger satisfaction because of the long wait times. The ability to predict taxi and Uber demand can help address the taxi-service inefficiency problem. With the deployment of the networked sensors and the widely used mobile phones in taxis, large amounts of information including location, time, number of passengers, weather, traffic, etc. can be collected in real time. This information provides opportunities to build an intelligent transportation system that is able to control and coordinate taxis at large scale and bring benefits to both taxi drivers and companies: taxi drivers can drive to high taxi demand areas, and ride-sharing companies (e.g., Uber) may re-allocate their vehicles using the surge pricing in advance to meet the passenger demand. Recent studies have shown that the passenger demand information can be used in the taxi dispatch system to reduce passenger waiting time, taxi cruising time, or supply re-balancing cost [2], [3], [4]. In this paper we only study the taxi and Uber demand prediction problem. The design of the taxi dispatch system has been widely studied in the transportation and the operation research area and is not considered here. E.g., it has been found that with the taxi demand prediction component, the taxi dispatch system can reduce the average total taxi idle distance by 52%, and the supply demand error by 45% [5]. We define the taxi demand prediction problem as follows: given historical taxi demand data in a region, we want to predict the number of taxis that will emerge within the next time interval. Inspired by previous works [6], [2], [7], [8], [9], [10], we aim to predict the met taxi demand. We use the number of pick-ups as a representation of the taxi demand in a region and treat them as time series data (see Fig. 1). Our method is general and can also be applied to predict the unmet taxi demand. As we discuss in Section 7.2, unmet demand can be inferred from the met taxi demand [11], [12]. In a recent report by the Taxi and Limousine Commission (TLC) [13], the agency that is responsible for regulating for hire vehicles in New York City (NYC), there is a strong correlation of the socioeconomic impact to the taxi demand at various governing zones such as building blocks. As such, we elect to study our technique at the building block level. Many methods have been proposed to predict taxi demand, including uncertainty analysis [6], probabilistic models [9], time series (ARIMA) [7], [8], SVM [2], and deep learning (LSTM) [14]. However, to apply these methods, we must answer two key questions. The maximum predictability captures the degree of temporal correlation of the taxi demand time series by measuring the regularity of human mobility. For most regions, the taxi demand is governed by a certain amount of randomness (e.g. irregular events, basketball match) and some degree of regularity (e.g. weekly patterns, peak during 4- 5pm weekdays), which can be exploited for prediction. For example, a building block with $\Pi^{max} = 0.8$ indicates that for about 20% of the time the taxi demand of this block appears to be totally random. In other words, no matter how good the predictive algorithm is, we cannot forecast the future taxi demand for a building block with

ΠImax = 0.8 with an accuracy that is higher than 80%. ΠImax represents the fundamental limit for predictability of the taxi demand. The ΠImax is a value between 0 and 1, the higher the more regular will the taxi demand be. Previous work assumed that the maximum predictability (the degree of the temporal correlation) in different regions is the same, and proposed the use of a single predictive algorithm for all regions [7]. However, the strong temporal correlation of taxi demand does not always hold. Different regions have different functions and thus different predictabilities (see Section 5.2). Fig. 1 shows the hourly taxi demand for two building blocks in NYC. The taxi demand near the Metropolitan Museum of Art (MoMA) (Fig. 1 top) exhibits a strong temporal pattern. The regular peaks in MoMA happen during the weekends (especially after the closing time): people usually visit museums during weekends and leave after the closing time. In contrast, the taxi demand near the west port (Fig. 1 below) appears to be more random. There is no clear temporal pattern near the west port. This is because the taxi demand in a transportation hub such as the west port is heavily dependent on the arrival of ships and there is a high variability in their arrival times [16]. In fact, the taxi demand near MoMA has one of the highest predictabilities among all the building blocks in NYC, and the taxi demand near the west port has one of the lowest. Intuitively we should use different predictors forecasting the taxi demand in these two building blocks. For MoMA it is better to use a predictor that is able to capture the temporal correlation, for example, a Markov predictor. For the west port, a predictor that uses machine learning and can capture exogenous features such as the ship schedule may be more effective. We posit that to select the best predictor, we must analyse the maximum predictability (ΠImax) of the taxi demand in each region.

1.1 What is Data Engineering?

The key to understanding what data engineering lies in the "engineering" part. Engineers design and build things. "Data" engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state. These pipelines must take data from many disparate sources and collect them into a single warehouse that represents the data uniformly as a single source of truth.

1.2 How Did Data Engineering Come About?

Many would say that data engineering as a profession has been around for well over a decade, maybe a couple, ever since databases, Microsoft SQL Servers and ETL came to be. Some would say ever since IBM popularized database management systems in the 1970s. With that, here's a very brief history recap.

In the 1980s the term "information engineering" was coined to largely describe database design and to include software engineering in data analysis. Somewhere after the rise of the internet in the 1990s and 2000s, 'big data" came to be. Yet DBAs, SQL Developers and IT professionals working in the field were not labelled "Data Engineers" at that time.

1.3 Why the Critical Need for Data Engineering Now?

By now you've heard/read about Gartner's determination back in 2017 that 85% of big data projects fail. This was largely due to a lack of reliable data infrastructures. Data could not be trusted enough to base key business decisions on it. Fast forward to 2019 and things had not improved. The CTO of IBM said that 87% of data science projects never make it into production. Gartner reiterated its prediction that now just 80% of projects would fail. A New Vantage Report produced similar stats.

Over the last decade, most companies have completed a digital transformation. This has produced unimaginable volumes of new types of data and much more complicated data at a higher frequency. While it was previously apparent that Data Scientists were needed to make sense of it all, it was less apparent that someone needs to organize and ensure this data's quality, security, and availability for the Data Scientists to do their jobs.

So, in the early days of big data analytics, Data Scientists were very often expected to build the necessary infrastructure and data pipelines to do their work. This was not necessarily in their skill sets or expectations for the job. The result was that data modelling would not be done correctly. There would be redundant work and inconsistency in the use of data among Data Scientists. These kinds of issues prevented companies from being able to extract optimal value from their data projects, so they failed. It also led to a high rate of Data Scientist turnover that still exists today.

Today with the onslaught of completed corporate digital transformations, the Internet of Things and the race to become AI-driven, it is crystal clear that companies need Data Engineers in abundance to provide the foundation for successful data science initiatives.

This is why will we continue to see the role of Data Engineers grow in importance and breadth. Companies need teams of people whose sole focus is to process data in a way that allows them to extract value from it.

1.4 What is the Relationship and Difference Between Data Scientists and Data Engineers?

Much has been written about the relationships between these two roles, so we'll be brief. In the past, companies thought that they could get away with having Data Scientists do the role of Data Engineers. This is what has caused much of the "unicorn effect" and shortage in Data Scientist recruitment.

Some Data Scientists also sold themselves as being able to do a Data Engineer's job. Many fell short – see the image to the right courtesy of O'Reilly.com.

Today, the volume and speed of data have driven Data Scientist and Data Engineer to become two separate and distinct roles albeit but with some overlap.

It's now widely recognized that companies need both Data Scientists and Data Engineers in an advanced analytics team. It's difficult to do any meaningful data science without Data Engineers to support this function. There's frequent collaboration between Data Engineers and Data Scientists however the priority skills and knowledge of tools are different.

Data Engineer Ability:

Data Scientists are focused on advanced analytics of data that is generated and stored in a company's databases. Data Engineers design, manage and optimize the flow of data with those databases throughout the organization. So, Data Scientists will be highly skilled in math and statistics, R, algorithms and machine learning techniques. Data Engineers will be more versed in SQL, MySQL, and NoSQL, architecture and cloud technologies and frameworks such as agile and scrum.

Both will likely know Python, visualization techniques and have other coding languages in common.

Foundation software engineering – Agile, devOps, architecture design, service-oriented architecture.

Distributed systems – This would include software engineer skills and software architect skills.

Open Frameworks – Apache Spark, Hadoop, perhaps Hive, MapReduce, Kafka and others…

SQL – This is a database staple and remains that way.

Programming – Python has become the favoured language for working with data. Java on the other hand, while still widely sought has fallen out of favour with most data scientists and engineers. Scala is another language that Apache Spark and Kafka are based on.

Pandas – a Python library for cleaning and manipulating data.

Visualization/dashboards

Cloud platforms – AWS is probably the most prevalent cloud skill set for Data Engineers to know. Google Cloud Data Engineering and Microsoft Azure are right behind.

Analytics – While mainly the realm of data scientists, statistical analysis skills or understanding of some of the different mathematical principles or probabilistic principles are necessary for being able to properly manipulate the data so that it is in a shape that is accessible for the people who are doing the end analysis on it.

Data modeling – Data modeling knowledge is quite important now in the sense that a Data Engineer needs to know how they are going to structure tables, partitions, where to normalize and deformalize data in the warehouse, etc. and how to think about retrieving certain attributes.

II. LITERATURE SURVEY

Title: Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions

Author: Hamparsum Bozdogan

Abstract: During the last fifteen years, Akaike's entropy-based Information Criterion (AIC) has had a fundamental impact in statistical model evaluation problems. This paper studies the general theory of the AIC procedure and provides its analytical extensions in two ways without violating Akaike's main principles. These extensions make AIC asymptotically consistent and penalize overparameterization more stringently to pick only the simplest of the true models. These selection criteria are called CAIC and CAICF. Asymptotic properties of AIC and its extensions are investigated, and empirical performances of these

criteria are studied in choosing the correct degree of a polynomial model in two different Monte Carlo experiments under different conditions.

Title: Limits of pre-dictability in human mobility

Author: Albert-László Barabási1

Abstract: A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals, raising a fundamental question: To what degree is human behavior predictable? Here we explore the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users. By measuring the entropy of each individual's trajectory, we find a 93% potential predictability in user mobility across the whole user base. Despite the significant differences in the travel patterns, we find a remarkable lack of variability in predictability, which is largely independent of the distance users cover on a regular basis.

Title: Data-driven robust taxi dispatch under demand uncertainties

Author: George J. Pappas

Abstract: In modern taxi networks, large amounts of taxi occupancy status and location data are collected from networked in-vehicle sensors in real-time. They provide knowledge of system models on passenger demand and mobility patterns for efficient taxi dispatch and coordination strategies. Such approaches face new challenges: how to deal with uncertainties of predicted customer demand while fulfilling the system's performance requirements, including minimizing taxis' total idle mileage and maintaining service fairness across the whole city; how to formulate a computationally tractable problem. To address this problem, we develop a data-driven robust taxi dispatch framework to consider spatial-temporally correlated demand uncertainties. The robust vehicle dispatch problem we formulate is concave in the uncertain demand and convex in the decision variables. Uncertainty sets of random demand vectors are constructed from data based on theories in hypothesis testing, and provide a desired probabilistic guarantee level for the performance of robust taxi dispatch solutions. We prove equivalent computationally tractable forms of the robust dispatch problem using the minimax theorem and strong duality. Evaluations on four years of taxi trip data for New York City show that by selecting a probabilistic guarantee level at 75%, the average demand–supply ratio error is reduced by 31.7%, and the average total idle driving distance is reduced by 10.13% or about 20 million miles annually, compared with non-robust dispatch solutions.

Title: Discovering regions of different functions in a city using human mobility and POIs

Author: Xing Xie

Abstract: The development of a city gradually fosters different functional regions, such as educational areas and business districts. In this paper, we propose a framework (titled DRoF) that Discovers Regions of different Functions in a city using both human mobility among regions and points of interests (POIs) located in a region. Specifically, we segment a city into disjointed regions according to major roads, such as highways and urban express ways. We infer the functions of each region using a topic-based inference model, which regards a region as a document, a function as a topic, categories of POIs (e.g., restaurants and shopping malls) as metadata (like authors, affiliations, and key words), and human mobility patterns (when people reach/leave a region and where people come from and leave for) as words. As a result, a region is represented by a distribution of functions, and a function is featured by a distribution of mobility patterns. We further identify the intensity of each function in different locations. The results generated by our framework can benefit a variety of applications, including urban planning, location choosing for a business, and social recommendations. We evaluated our method using large-scale and real-world datasets, consisting of two POI datasets of Beijing (in 2010 and 2011) and two 3-month GPS trajectory datasets (representing human mobility) generated by over 12,000 taxicabs in Beijing in 2010 and 2011 respectively. The results justify the advantages of our approach over baseline methods solely using POIs or human mobility.

Title: Automatic city region analysis for urban routing

Author: Sasu Tarkoma

Abstract: There are different functional regions in cities such as tourist attractions, shopping centres, workplaces and residential places. The human mobility patterns for different functional regions are different, e.g., people usually go to work during daytime on weekdays, and visit shopping centres after work. In this paper, we analyse urban human mobility patterns and infer the functions of the regions in three cities. The analysis is based on three large taxi GPS datasets in Rome, San

Francisco and Beijing containing 21 million, 11 million and 17 million GPS points respectively. We categorized the city regions into four kinds of places, work-places, entertainment places, residential places and other places. First, we provide a new quad-tree region division method based on the taxi visits. Second, we use the association rule to infer the functional regions in these three cities according to temporal human mobility patterns. Third, we show that these identified functional regions can help us delivering data in network applications, such as urban Delay Tolerant Networks (DTNs), more efficiently. The new functional-regions-based DTNs algorithm achieves up to 183% improvement in terms of delivery ratio.

III. EXISTING SYSTEM:

Current, studies have shown that the passenger demand information can be used in the taxi dispatch system to reduce passenger waiting time, taxi cruising time, or supply re-balancing cost. With the deployment of the networked sensors and the widely used mobile phones in taxis, large amounts of information including location, time, number of passengers, weather, traffic, etc. can be collected in real time. The latest works use the deep neural networks combined with multiple features to predict the taxi demand in cities. Wang et. al [38] propose TaxiRec, a framework for evaluating and discovering the potential passengers of road clusters. TaxiRec includes three influential factors points-of interest, road length and road type in their neural network predictive algorithm. Wang et. al [39] use a deep neural network structure to discover complicated taxi supply-demand patterns. They utilize multiple data sources including car hailing orders, weather and traffic data. Different to these two papers, in this paper predict the taxi demand of each building block and we consider the taxi demand of one location as a time-series data. For one single building block, the road type, road length and point-of-interest types are fixed parameters and thus cannot be considered as features for prediction here. By showing the predictability of different land use (Table 3), we show that the predictability is correlated with the point of interests, and thus corresponds to the prediction accuracy of different predictors.

Disadvantage:

- The performance of this encrypted format is low.
- It is consuming more time and the cost.
- The security and the accuracy are less

IV. PROPOSED SYSTEM

The predictability can help determine which predictor to use in terms of the accuracy and computational costs. We quantify the limits of predictability of a location's taxi demand based on its taxi demand history. The latest works use the deep neural networks combined with multiple features to predict the taxi demand in cities. The temporal pattern of human mobility also leads to the temporal pattern of taxi pick-ups. The human mobility patterns for different functional regions are different. We show that the maximum predictability of the taxi demand can reach up to 83% on average capturing the degree of the temporal correlation of a taxi demand time series. Our findings indicate that taxi demand in NYC incorporates strong temporal patterns. We also find that the Uber demand is easier to be predicted compared the taxi demand due to different cruising strategies as the former is demand driven with higher temporal regularity. • We implement and compare the prediction accuracy of five commonly used and representative predictors and examine their performance under different maximum predictability: the Markov (probability based) [17], the Lempel-Ziv-Welch (LZW) (sequence modeling) [18], the auto-regressive integrated moving average (ARIMA) model (time series forecasting) [19], the Neural Network (NN) (machine learning) [10], and the Long Short-Term Memory (LSTM) (deep learning) [20]. Our results indicate that the maximum predictability is an approachable target for the actual prediction accuracy. • We observe that the LSTM predictors provides better accuracy for building blocks with low predictability ($\Pi^{max} < 0.83$) by capturing hidden long-term temporal dependency, while the Markov predictor provides better accuracy for building blocks with high predictability ($\Pi^{max} > 0.83$). A compute-intensive deep learning predictor does not always outperform a simpler Markov predictor, while the latter requires only 0.02% computation time. Knowledge of the predictability can help determine which predictor to use in terms of the accuracy and computational costs. Using a third time-series data set, we show that our finding is general and can be applied to other timeseries data sets.

Advantages:

- High security and more effective.
- This system gives more accuracy compared to another systems

Algorithm

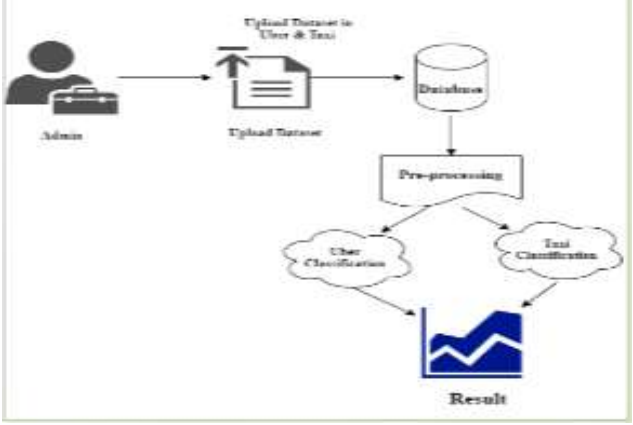
Ciphertext:

Ciphertext is also known as encrypted or encoded information because it contains a form of the original plaintext that is unreadable by a human or computer without the proper cipher to decrypt it. Decryption, the inverse of encryption, is the process of turning ciphertext into readable plaintext.

AES:

The algorithm described by AES is a symmetric-key algorithm, meaning the same key is used for both encrypting and decrypting the data. The Advanced Encryption Standard (AES) is a symmetric block cipher chosen by the U.S. government to protect classified information. AES is implemented in software and hardware throughout the world to encrypt sensitive data. It is essential for government computer security, cybersecurity and electronic data protection. Symmetric, also known as secret key, ciphers use the same key for encrypting and decrypting, so the sender and the receiver must both know -- and use -- the same secret key. The government classifies information in three categories: Confidential, Secret or Top Secret. All key lengths can be used to protect the Confidential and Secret level. Top Secret information requires either 192- or 256-bit key lengths.

V. SYSTEM ARCHITECTURE



VI. CONCLUSION

We find that there is a high predictability of taxi demand (up to 83% in average), which indicates strong temporal correlation of human mobility. We also examine which commonly used predictive algorithm could approach the maximum predictability. We show that the compute-intensive deep learning predictor does not always have better prediction accuracy than the Markov one. In the areas with low predictability ($\Pi max < 0.83$), the LSTM predictor can reach high accuracy by capturing the hidden long-term dependent temporal patterns. On the other hand, in the areas with high predictability ($\Pi max > 0.83$), the Markov predictor can reach high prediction accuracy while keeping the computation time low be better captured in the Uber taxi data, possibly due to different cruising strategies.

REFERENCES

[1] Y. Huang and J. W. Powell, "Detecting regions of disequilibrium in taxi services under uncertainty," in SIGSPATIAL'12, Redondo Beach, CA, USA, November 7-9, 2012, 2012, pp. 139–148.

[2] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset," in IEEE PerCom , Seattle, WA, USA, Workshop Proceedings, 2011, pp. 63–68.

[3] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases, ser. SSTD'11. Berlin, Heidelberg: SpringerVerlag, 2011, pp. 242–260.

[4] R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing-theoretical perspective," The International Journal of Robotics Research, vol. 35, no. 1-3, pp. 186–203, 2016.

[5] F. Miao, S. Han, S. Lin, J. A. Stankovic, D. Zhang, S. Munir, H. Huang, T. He, and G. J. Pappas, "Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach," IEEE Trans. Automation Science and Engineering, vol. 13, no. 2, pp. 463–478, 2016.

[6] F. Miao, S. Han, S. Lin, Q. Wang, J. A. Stankovic, A. Hendawi, D. Zhang, T. He, and G. J. Pappas, "Data-driven robust taxi dispatch under demand uncertainties," CoRR, vol. abs/1603.06263, 2016.