Commuter classification and behavior clustering: Beijing use case

Selene Baez Santamaria
s.baezsantamaria@student.vu.nl

Abstract

Public transportation, centered on subway and bus networks, is an data-rich domain that can benefit from data mining and machine learning techniques. The classification of commuters versus non-commuters/occasional travelers can help government, transport management and operators to better target their policies in order to improve the transportation network in large cities. Furthermore, characterizing commuters by behavior clustering can bring deeper insight into their needs and routines as a whole. This project proposes the usage of ensemble models for classification and clustering of public transport users. For this purpose, transit card data will be used, available from the city of Beijing, China.

Contents

1 Introduction

1.1 Urban public transportation

Urban public transportation includes systems that are available for use by anyone in urban areas. Its facilities are commonly composed by buses, subway/metro lines, light rails, tramways, trains, taxis and others. As a network, they provide service for the majority of citizens in urban areas.[25]

Figure 1 shows the passsenger transport usage, as million passengers per kilometer, in several different countries according to the Organisation for Economic Cooperation and Development (OECD). From all OECD countries, the United States, China, Germany, France, Italy, and the United Kingdom contitute the six countries with the most passenger transport, according to their reported data from 2015 or later.[17]

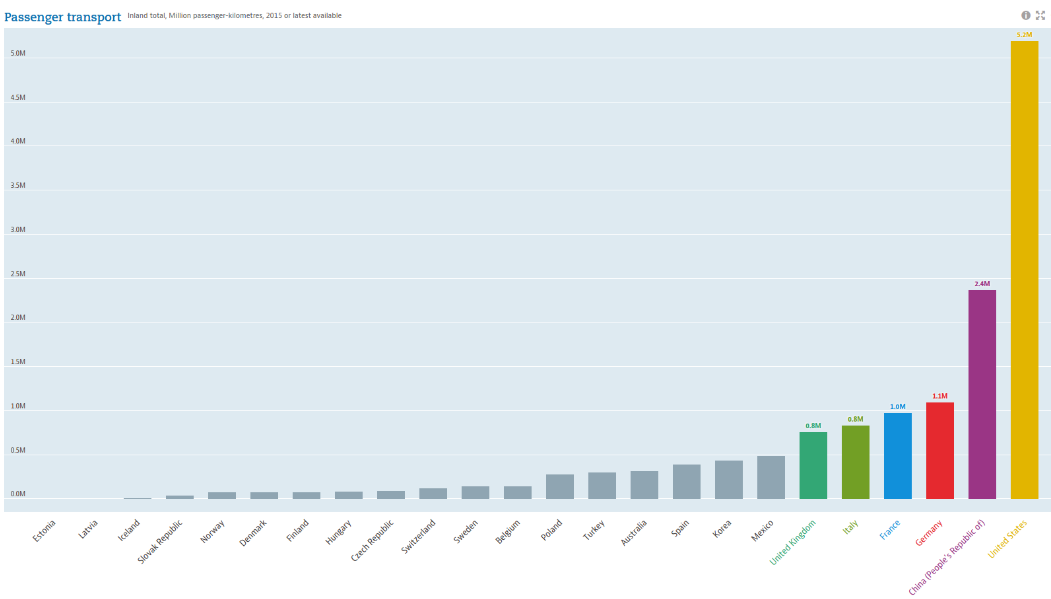


Figure 1: OECD countries and their passenger transportation data.

Furthermore, historical data in Figure 2 reveals the 15 years behavior for each of the aforementioned countries. Most of the countries show stability, with increase or decrease of less than 0.10 million passengers for European countries, and 0.5 million passengers for the United States. China, however, shows a trend with steep increase for most of the selected years. In fact, comparing to its less than 1.2 million passengers in 2000, China doubled its public transport usage to 2.4 million passengers in 2015.

Though it is a more sustainable alternative compared to private car usage, public transport usage has a significant environmental impact, affecting noise and air pollution. Diesel buses, which generally make up a major part of public buses, have large fuel consumption needs and contribute significantly to CO_2 emissions. Even eco-friendly alternatives such as hybrid diesel buses are sensitive to operating conditions, as their fuel consumption may increase by up tp 50% when the on-board air conditioning is on.[31]

Consequently, public transportation directly relates to energetic demand, since its facilities are mostly petroleum or electrical based. In terms of global energy consumption, passenger transportation accounts for about 25% of the total world energy consumption. Furthermore, the transportation sector consupmtion increases at an annual average rate of 1.4.% [9] This may bring further economical implications for countries with high public transportation demand.

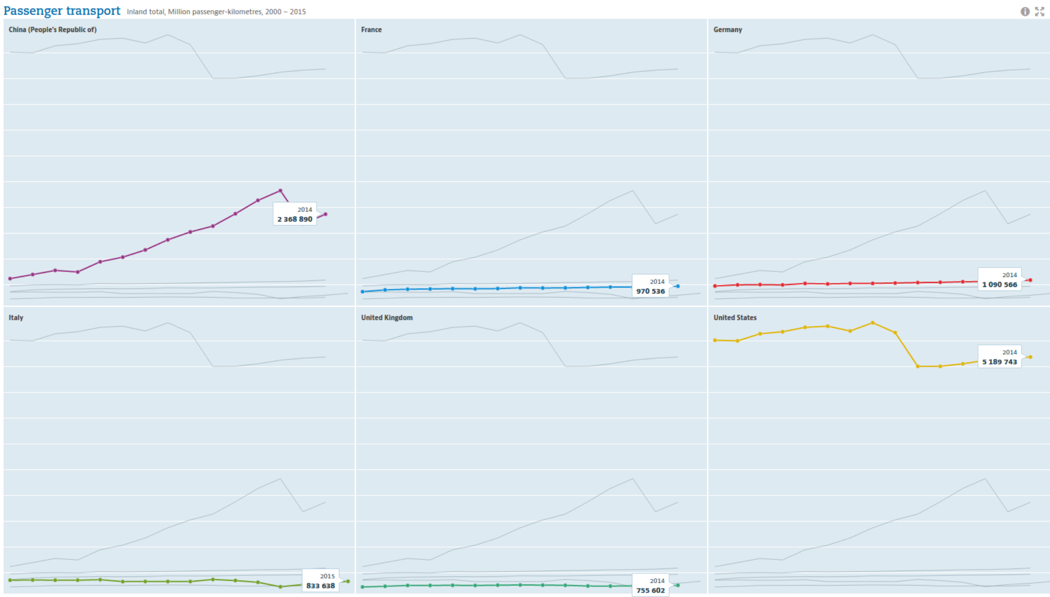


Figure 2: Historical data for the top six countries with most passenger transport usage. China (top-left image) has the steepest increase overall.

1.1.1 Who are the commuters?

A major proportion of public transport users is represented by commuters. These are regular users of public transit, with consistent spatiotemporal patterns in their travels. Driven by a routine, commuters travel back and forth from specific places, for example, from their home to work, school, or other similar locations.

As commuters are frequent users of public transit, the conditions of the public network directly influence their personal well being and generally impact their quality of life. Intuitively, if the commuting experience is unpleasant, daily travel can bring distress to commuters and/or even repel them from using the public transport at all. Several studies have looked into public transit evaluation from different perspectives, including commuters' needs [13]. The most common aspects of it include: travel time, average speed, delays, accessibility, service coverage, crowded level, facilities quality, and fare rate. Weng et al [27] identified five indexes (Convenience, Rapid, Reliability and Comfort) that summarize commuters priorities when choosing to travel by public transport.

From both of the above, the large presence of commuters and their known needs and preferences, it follows that identifying commuters and addressing their needs can help in creating a sustainable public transportation network. Public transit stakeholders should be able to understand the commuters' demands and its dynamics, consequently bringing long term planning and policies for improving the overall commuting experience.

1.2 The city of Beijing

The city of Beijing presents a special case of urbanization and rapid industrialization. This is reflected in a sudden population growth of 20% per decade since 1960, with the largest increase of 44% in the last ten years. The latest official census in 2010 reported the urban agglomeration of Beijing (including Beijing itself and its adjacent suburban areas) having a population of 19,612,368 people. The UN World Urbanization Prospects estimates the 2017 population at over 22 million inhabitants. [22]

As a result of the population explosion, many environmental and social resources are under pressure. From the environmental side, one of the most notable issues is related to air pollution, due to the significantly high pollutant emissions in the city [30]. Similarly, the city's downstream river pollution is serious, with most regions of the Yellow river being unable to comply with the lowest water quality standards. [26]

On the aspect of social resources, one of the main complications is mobility. In Beijing, public transport is the dominant mode of transportation, accounted for 44.0% of all trips compared to 32.6% attributed to private cars [13]. In 2008, the total ridership was 6.5 billion travels. Though the network is continually expanding, it is a fact that public transport is overcrowded, constantly reaching over 100% capacity [18].

Beijing public transport is composed of buses, subway and bicycles. The three types can be accessed by using a single smart card.

Bus: In 2015, there were 876 bus lines with 23,287 buses in operation. The bus network is the most extensive mode of transportation, expanding over 20,186 km. It observes an average daily traffic volume of 10.98 million passengers, with the highest daily volume reaching 13.07 million on one day. [3]

Subway: The Beijing subway has 18 lines with 334 stations, of which 53 are transfer stations. In 2015 it had an operating length of 554 km, with 5,024 vehicles running. [3] Its network is split by two operators: the state-owned Beijing Mass Transit Railway Operation Corp (operating 15 lines), and the joint Hong Kong venture Beijing MTR Corp (operating 3 lines).

Beijing's subway has an average daily traffic volume of 9.11 million passengers, with a maximum recorded volume of 11.66 million passengers. As such, it is the second busiest metro system in the world, providing 3,410 million annual journeys. Compared to the service provided in 2012, the system observed a 39% increase in usage by 2014. It is also the second longest metro network, surpassed by Shanghai by only 21 km. [24]

Bicycles: Beijing first implemented public bicycle systems in 2012. As of 2015, in total, 67,000 bikes are available for rental with 2,700 pick up/drop off points spread across the city. [3]

1.3 Smart cards and Big Data

Smart cards present us with a straightforward way of massively collecting daily data. In the last years, smart card systems have become more popular in the Transportation domain, making it possible to monitor travelers transactions and facilitating fare collection. Several cities have implemented such systems, for example the Octopus card in Hong Kong[5], Oyster card in London [2], OV-chipkaart in The Netherlands [6], and Yikatong card in Beijing [4], to name a few.

In Beijing, over 90% of public transit users are smart card holders. There is a significant incentive for using the Yikatong smart card since bus rides are heavily subsidized (users have only to pay 50% of the full price)[11]. Moreover, the Yikatong smart card system is also integrated with taxi, electricity and sewage payments, making it convenient to use as a general paying method.

Data quantity Placed in context, public transit systems serve at least hundreds of users daily, where a typical user performs several trips a day, every day. On the specific case of Beijing, there are hundreds of thousands of smart cards gathering between 5 and 16 million records (trips) a day, among a large complex network containing thousands of routes and tens of thousands of stops.

Data quality However, though smart cards exponentially increase the quantity of data, they do not completely guarantee its quality. As is, some aspects of the trips cannot always be faithfully recorded but are inferred (for example, the transfers between the subway system when no check-in/out is done at changing trains). Furthermore, some fields are sometimes simply missing or incorrectly recorded due to malfunctions and situations out of control.

Given the large amounts of data collected and its nature, the analysis of such becomes challenging. Transit smart cards are capable of recording spatiotemporal information at an individual level over long periods of time. This generates a large volume of historical data that only tailored big data techniques can deal with.

1.4 Project motivation

This project performs an interdisciplinary study between the areas of Artificial Intelligence and Metropolitan Transportation. It is focused on introducing data mining techniques to a data rich domain.

The area of Artificial Intelligence is able to provide dozens of prediction algorithms. Though constantly under refinement, it is time for state-of-the-art techniques to be applied, tested and validated under real and large impact situations to test their ability to deal with noisy streams. Comparably, given the ever growing complexity of urban mobility, domain experts must focus on analyzing trends and insights instead of curating and making sense out of raw data. As such, introducing these state-of-the-art techniques into the Metropolitan Transportation domain can aid to unravel massive human behaviors and reveal patterns and trends in mobility.

1.4.1 Societal context

Identifying and analyzing mobility patterns may have different goals, from description, prediction or prescription, all of which affect their stakeholders directly.

A descriptive analysis determines how people use the public transit. It can pinpoint chaotic hotspots in the city, peak hours, popular routes or other behaviors. A predictive analysis investigates how will people use the transit in the future or under new circumstances. For example, public transport usage projections in the years to come directly affects environmental models trying to improve air and water quality, energetic demand or other natural and economic resources.

Finally, a prescriptive analysis focuses on how should the different stakeholders deal with mobility behaviors. For example, the government as well as transport management and operators would gain invaluable spatial and temporal insights regarding commuters' behaviors. This insight may lead to tangible results, including policies for increasing the efficiency of the public transit network, adjustable travel fares tailored to the most relevant mobility patterns, incentives to relieve peak hours and thus traffic congestion, urban planning for residential and industrial land use, and others.

Given that Beijing has a widely spread data collection system, combined with formidable institutions capable of introducing new measures in their public transportation network, this city is an excellent use case where the results of an in-depth study can generate actionable plans and bring benefits in the short and long term.

Furthermore, the social context of Beijing presents specific opportunities for improvements where the full power of data analysis and its impact can be tested. For example, the city of Beijing faces a large imbalance between residential and working areas. Due to urban expansion, most residents have been forced to move to suburban areas due to the lack of affordable housing, regardless of having their work environments within the six Ring Roads [32]. Investigating and targeting this group could alleviate the pitfalls of long distance commuting.

1.4.2 Scientific context

Mobility patterns in metropolitan areas follow complex swarm behaviors. Based on individual travels and routines, travelers exhibit distinguishable characteristics on a larger scale. Both individual and collective levels of understanding are crucial for Transportation experts. In order to explore both levels, Metropolitan Transportation studies typically focus of the usage of surveys. These surveys are targeted to reach travelers on an individual level, while large scale indicators and aggregated data are taken to investigate their collective behavior.

These methods have several disadvantages. On the one hand, surveys are costly to implement, and in general have problems related to small non-representative samples. Even when these problems are escaped, the usual quality versus quantity trade off is present, reducing the confidence of the collected information. On the other hand, large scale measurements (i.e. total passenger flow) miss the interactions between individuals that cause the collective behavior.

On top of this, an important consideration on the Metropolitan Transportation domain is that the data collected by smart cards is unlabeled. This means that traveling behaviors are not assigned to known specific categories, making it hard to validate and evaluate. Typically, this issue is address by asking some sample users -via surveys- how they categorize themselves (for example, if they consider themselves to be commuters) and then extrapolating this profile to new users. However, self-reported data by itself has bias problems, therefore introducing noise or false patterns.

Fortunately, the field of pattern recognition has seen major development in the last years. Nowadays, there exist machine learning and other data mining methods specialized in analyzing disaggregated complex information. Data analysis can be as general is specialized as needed, producing reliable and comprehensible information and visualizations. Furthermore, unsupervised tools have arisen that find patterns based on the data alone, thus being independent from the aforementioned biases.

1.5 Thesis organization

The rest of this thesis is organized as follows: On the next section we perform a *Literature review* to explore previous work on mining smart card transit data. We also summarize current representation and pattern recognition methodologies for dealing with complex spatiotemporal data.

Subsequently, we establish the *Research framework* where we explicitly state the objectives and research questions of this project. As a result, we limit the project's scope and clearly define the most important terms to be used.

We continue to describe the *Methodology* thoroughly. This consists of an extensive description of the data and its characteristics, our proposed 3 dimensional representation for spatiotemporal data, and the data mining approach to follow, including supervised and unsupervised learning techniques for dimensionality reduction and pattern recognition.

Following this, we identify three distinct stages of the project: *Data preparation and preprocessing*, *Commuters identification*, and *Traveling behavior clustering*. In the *Data preparation and preprocessing* section we describe the pipeline for processing raw data, extracting trip attributes and finally creating the proposed 3D representation of an user's traveling behavior.

The section on *Commuters identification* describes a supervised learning approach for classifying labeled data, using feature selection and ensemble models. Its counterpart, the section on *Traveling behavior clustering* describes an unsupervised learning approach to recognize similar traveling behaviors, using feature extraction (by means of an autoencoder) and clustering algorithms.

Finally, we gather conclusions regarding the proposed representation. We compare both supervised and unsupervised approaches and explore future work opportunities.

2 Literature review

In this section we look at studies within the last decade that are related to smart card transit data. First, we summarize the approach and the most relevant findings of each paper in order grasp a broad view of the Transportation domain. Secondly, we explore representations for spatiotemporal data and compare the way traveling behavior is usually represented in the Transportation domain, and other types of representations available in the Artificial Intelligence domain. Thirdly, we discuss techniques for pattern recognition through supervised and unsupervised learning. Finally we explore some pioneer work in end-to-end learning.

2.1 Data mining on transit card data

With the introduction of smart card systems in large cities, several studies have aimed to extract knowledge from the large amounts of data collected. Many of this studies focus on analyzing traveling behavior, which is regarded as a spatiotemporal mobility pattern. Though different in their methodology, results concerning commuters are duplicated across studies. As the spatial and temporal regularity of commuters' travel behavior is evident in their smart card data, they pose an excellent opportunity of study.

Morency et al. study spatio-temporal variability in Canadian smart card data. On the one hand, they examine spatial variability by measuring the number of distinct stops a smart card user visits, and the frequency of each stop. On the other hand, they examine temporal variability by clustering the boarding times of each type of smart card. Using these features, they observe the week to week variability for each of the five types of transit card available (Adult-interzone, Adult-express, Adult-regular, Elderly and Student). Their findings show that commuter types of cards visit a smaller range of bus stops compared to non-commuter types. Therefore, a small number of stops account for a high proportion of commuter's boardings. Additionally, commuters have the highest proportion of zero-boarding days on weekends [15].

Bhaskar et al. are concerned with passenger segmentation using Australian smart card data. First, they perform a two level DBSCAN algorithm for investigating spatial patterns, where the first level clusters Destination stops and the second level clusters Origin stops. From this they extract frequent Origin-Destination (O-D) pairs. Separately, they applied DBSCAN to temporal features to determine most frequent boarding times. As such, they characterize each user by the percentage of journeys they perform between the regular O-D, and the percentage of journeys they perform during their habitual times. Users with at least 50% spatial and temporal regularity are thus classified as transit commuters; while users with no evident spatial or temporal pattern are classified as irregular passengers. The authors find that while most (64%) of the passengers riding the public transit are irregular passengers, it is transit commuters who bring the most (46%) revenue. Furthermore, they find that irregular passengers prefer high frequency routes significantly more than transit commuters, arguing that commuters are usually on a time habit, and thus are more willing to check and adapt to public transit timetables. [1]

Tu et al. follow a supervised learning approach to classify public transit users in Beijing as commuters or non-commuters. In order to produce labeled data, they convey an online survey asking for travel patterns and smart card ID. Matching the ID to the journeys recorded by smart card during the span of one week, they collect records associated to 978 travelers. The classification is then performed by a Support Vector Machine (SVM), which reaches up to 94.24% accuracy. [23]

Langlois et al. present an innovative representation for smart card data. Using four weeks worth of data from London Oyster cards, they represent the card information as a time-ordered sequence of inferred activities. 11 clusters are found and characterized by evaluating socio-demographic variables like age, employment, annual household income, children per household and vehicles per household. The authors further grouped the clusters under "working day", "home bound", "complex activity pattern" and "interrupted pattern" categories. Their findings show that four clusters, grouped under the "working day" category have significantly different activities during weekdays as compared to weekends, with some avoiding transit during the weekends and others visiting different areas. Four more clusters, grouped under the "home bound" category, are characterized by staying mostly at their primary area and low number of traveled days. [10]

One of the latest work on the field corresponds to Ma et Al. The objective of their work is to determine a scoring function for travelers that can correctly identify them as commuters, or non-commuters. In their work, they cluster stops using an improved DBSCAN algorithm. They engineer features for representing the frequency in which travelers follow spatio-temporal patterns. Travelers are then clustered according to these features following the ISODATA algorithm. As an output of the clustering, optimal cutoff levels in the scoring function were determined. As a result, evaluating a traveler does not depend on clustering centroids, but only on calculating the commuting score. This, as expressed by the authors, reduces computing time and treats each traveler independently from the others, which is not true for clustering algorithms [11].

A common practice, as used by [11], [10], and [15] is to divide the day into -hourly or half-and-hour-time bins. Bhaskar et al. recognize this as a problem in the field, by pointing out that this design choice segregates journeys from 9:59 AM and 10:01 AM even though they intuitively belong to the same behavior.

2.1.1 Volume of data

The volume of data collected by smart card systems is massive and is usually impossible to analyze all of it at once. The volume of the samples analyzed by previous work ranges from hundreds of smart cards to tens of millions of smart cards, leading to up to hundreds of millions of individual smart card transactions. The details of the revised literature are summarized in Table 1.

Authors	Year of publication	Records	Unique smart cards	Time span
Morency et al. [15]	2007	2.2 million	7,118	277 days
Ma et al. [12]	2013	Unknown	3 million	one week
Ortega [19]	2013	65 million	5.7 million	one week
Bhaskar et al. [1]	2015	34.8 million	1 million	4 months
Tu et al. [23]	2016	8,067	978	one week
Langlois et al. [10]	2016	3 million	33,026	four weeks
Ma et al. [11]	2017	364 million	18 million	one month

Table 1: Volume of data analyzed by different authors

Given the limit on how many records can be examined per study, researchers usually face the decision to reduce the dataset to a manageable size. As such, there exists a trade off between the number of unique smart cards and the time span of the collected data. Some researchers, like Ortega [19], decide to analyze a large population over short periods of time. Others, like Bhaskar et al. [1] choose to explore long term behavior thus having to reduce the population size.

However, it is worth noting that the total number of records studied has increased overtime. This most likely is due to the trends of increased computational power and the design of optimized mining algorithms. A clear example is the study by Ma et al. [11] published just this year that was able to include data of a significantly large population over a month.

2.2 Representing spatiotemporal data

As Marr puts it, representations make explicit different types of information implicit in entities [14]. Thus, representations mainly differ in the information they describe and the way they describe it. Usually, representations are generated to achieve a information processing goal. Thus, the value of a representation depends on the purpose of the task it will be used for.

Data representation is one of the fundamentals in data mining. Ideally, the representation of a data point is comprehensive of its underlying unique factors and leaves out unnecessary or noisy information. Furthermore, the format for the representation must be akin to the types of information that data mining algorithms can process. Therefore, finding suitable representations for complex concepts like space and time is not an easy task.

2.2.1 Traditional feature engineering

Human mobility is intrinsically tied to spatio-temporal properties. Still, the greatest amount of studies analyze public transit journeys by separating spatial features from temporal features. Furthermore, in general scalar aggregated features are used for users characterization. Some examples are:

- **Frequency indicators:** number of traveled days [1] [10] [11], number of journeys [1], number of times a stop was visited [15], number of days with zero boardings [15], most frequent home/work stop [11], most frequent home/work route [11], most frequent departure time from home/work [11], number of trips to the most frequent home/work stop[11], number of trips following the most frequent home/work route [11], number of trips during most frequent departure time from home/work [11]
- **Range/coverage indicators:** distinct stops visited [15], spread of days between the first and last journey [10]
- **Calendar-based indicators:** observed day [15], day of week [15]

Though popular among the Transportation domain, hand engineered features may present great disadvantages. While these features are intuitive and semantically meaningful for Transportation specialists, they do not always represent distinctive properties of users or their public transit journeys. Therefore, the time invested in designing and producing features may not always payback in relevant findings.

This case can be compared with the trends seen in Computer Vision. A few decades ago, most approaches for Image Understanding were focusing on designing features to describe them (i.e. SIFT). However, after the rapid development of Neural Networks in the last years, the most successful Vision applications are based on learned features. As Nithin and Sivakumar explain, hand crafted features are time consuming, fragile and incomplete, thus being outperformed by automatically extracted features which learn better the underlying representations in images [16].

2.2.2 Feature extraction

Feature extraction refers to the creation of features that represent the underlying characteristics of data. These features are automatically created, solely from the data, in either statistical or learned ways. One popular way for extracting features is by using methods for dimensionality reduction, which beyond finding representations further tackles the curse of dimensionality.

Principal Component Analysis

One of the most robust algorithms for this is Principal Component Analysis (PCA) which is a mathematical tool used across several domains. By doing matrix manipulation, PCA extracts eigenvalues and eigenvectors from a given dataset. The top eigenvectors represent the ways in which the data points are more different from each other.

An isolated work related to this was performed by Langlois et al. Following a unique methodology for engineering features, first they represent the travel data per user using a three dimensional matrix where x represents the day in the four week period, y represents the hourly time bin, and z represents the area where the inferred activity took place, encoded as a one hot vector. The authors perform PCA for dimensionality reduction, based on Eagle and Pentland's eigenbehaviours [8]. An analysis of the average correlation of the first 13 components, results in the selection of the first 8 components as the most informative and stable. The projections of a user sequence onto these components (called weights) constitute the features to be clustered using k-means. [10]

Autoencoders

Following the same main principle as the first neural networks, autoencoders are highly connected networks that map high dimensional data to low dimensional spaces. Primarily used for images, the goal of an autoencoder is to deconstruct an input image onto a representation, and reconstructing the image again, with a minimum loss of information. Each of these are called the Encoder and Decoder modules, respectively. Together these are learning modules that tune its parameters until achieving sufficient performance.

Much research has been done on autoencoders, leading to several variations of them. Denoising autoencoders result in a more robust algorithm, since they get a corrupted image as input, but aim for reconstructing the original image. Therefore, the autoencoder does not simple map one instance to a representation, but truly learns the significant characteristics present in the data [16].

To the best of our knowledge, these techniques have not been introduced to the Transportation domain.

2.3 Pattern recognition on spatiotemporal data

2.3.1 Classifying algorithms

The domain of Metropolitan transportation faces a specific problem: although smart card systems have allowed massive collection of data, this data is not labeled regarding commuting behaviors. Additionally, obtaining labels for smart card data is expensive and unreliable, since it has to be acquired through surveys or interviews. Furthermore, even when labels are obtained, the amount of labels obtained is often insufficient for big data analysis. It is due to these reasons, that most studies are inclined to used unsupervised learning techniques.

One of the few studies that uses labeled data corresponds to Tu et al. They obtain 978 labeled records, with an almost equal distribution of records over both classes (49.18% related to commuter samples and 51.82% related to non-commuter samples). They solve the issue of limited samples by selecting a model that is not heavily affected by sample size: Support Vector Machines. Their results report a 94.24% accuracy over a test set of 295 samples.

2.3.2 Clustering algorithms

If labeled data is not available, then unsupervised learning techniques must be applied. There is a large variety of clustering algorithms available nowadays, however not all of them are suitable for all types of data and purposes.

Hierarchical clustering

Langlois et al. use agglomerative hierarchical clustering for areas clustering. In order to infer the user-specific activities, all stops or stations visited by each user are clustered by merging the two closest areas until a threshold distance is reached. Their algorithm also considers the distance between stops and the frequency of travel between them. Therefore, different activities are likely to be associated with different areas [10].

Partitional clustering

The K-means algorithm is the most widely used method for partitional clustering. It requires having a predefined number of clusters to fit the data to.

Morency et al. use K-means for clustering hourly boarding times according to card type. They apply Hamming distance (representing the percentage of data between two elements) and a combination of batch and online updates. Through empirical tuning, they select to find four clusters per card type. It is worth noting that by using a card-day unit, they allow a card to belong to a different cluster according to the day of travel. As every card type is composed of four boarding patterns, travelers are not restricted to follow a routine everyday, but can exhibit different behaviors on different days. For example, the Adult-regular card type contains a 9:00AM-and-5:00PM-boarding cluster and a no-boarding cluster. Thus, a user of this card could belong to the first cluster on weekdays and to the second cluster on weekends. [15]

Bhaskar et al. apply K-means for binary classification purposes. As such, they classify frequent and infrequent transit users, using the number of traveled days and the number of journeys made as features. Unfortunately, K-means performs poorly since no distinct clusters are evident. The most likely cause for the previous is the strong correlation between traveled days and journeys, combined with the authors oversight of whitening and standardization techniques. [1]

Langlois et al. use K-means to find clusters of activity sequences. They employ specialized sampling techniques, like bootstrapping, to deal with big data. Moreover, they tune the algorithm parameters using the DB-index, which is the ratio of the within cluster distances to the across cluster distances. They find two optimal number of clusters (4 and 11), out of which they select the largest to provide the most detailed segmentation. They further perfection the algorithm by using k-means++initialization over 150 replications. Additionally, this paper acknowledges that clustering techniques are sampled based, which means different samples may find different optimal solutions. The authors validate their approach by analyzing the stability of the clusters over samples obtained at different points in time. By extracting the same number of clusters and fitting the samples to each set, they find that 91% of users are assigned to their equivalent clusters. [10]

Density based clustering

Density based algorithms excel at dealing with anomalies, since they ignore low density areas and interpret them as noise. They do not required a redefined number of clusters and adapt to find clusters of any size. The required parameters for DBSCAN are a maximum reach distance ϵ and the minimum number of points per cluster.

Bhaskar et al. use three DBSCAN algorithms to cluster Origin stops, Destination stops, and boarding times. For each of the previous, they tune the algorithm parameters by fixing a domain reasonable ϵ (1000 m walking distance or 5 min variance in boarding time), and selecting the minimum points by comparing the percentage of data considered to belong to any cluster as opposed to data considered to be noise given the par-specific parameters [1].

Ma et al. use an improved DBSCAN algorithm to cluster bus/subway stops. In their approach, abnormal stops are not considered noise, but are allowed to be re-clustered by splitting large clusters into several smaller clusters.

Though clustering algorithms are common in the field, they are not always used for classifying users. For example, Bhaskar et al. use density based clustering for engineering regularity features. However, the classification of users is rule-based according to which feature (spatial or temporal regularity) is stronger in each user [1]. Morency et al. use partitioning clustering to characterize existing user categories according to their boarding times [15].

As a conclusion, we note that while there has been research applying basic clustering and classification algorithms, most studies lack further specialized data mining techniques for preprocessing data, tuning algorithms parameters, and/or visualizing results.

2.4 End to end learning

As mentioned in Section 2.2, the effectiveness of a representation depends on the task for which it is used. Hence, it is reasonable to believe that if the representations are able to learn according to the task performance, then the representations will improve leading to an improvement in the task itself. This is an example of end to end learning, a principle applied in new popular areas such as Deep Learning for image classification.

Though applied to supervised learning tasks, end to end learning is yet to be studied for unsupervised tasks. While supervised learning has quantitative ways of evaluation (such as objective functions to be minimized), unsupervised learning cannot be evaluated in the same way due to the lack of labels. taking our case for pattern recognition tasks, the error in classification can easily be identified whereas there is no intuitive error to measure for clustering.

There exists a few pioneer studies that explore this idea.

2017 Depict [7]

2016 Jule [29]

2016 DEC [28]

3 Research framework

The underlying goal of this project is to find an accurate spatiotemporal representation for public traveling behavior while accounting for big data constraints and the inherent data nature. The main two objectives are:

Objective 1 To identify commuters based on their routine patterns.

Objective 2 To group users with similar travelling behaviors.

Combined, these objectives characterize public transit users in the city of Beijing. For this project we decide to use one month worth of smart card data, since we believe it to be a long enough period to see different traveling patterns, while keeping the data at a manageable level.

3.1 Research questions

The main objectives is further broken down into answering the following research questions:

1. How can spatiotemporal features be analyzed as a unit?
2. What are the most relevant features when identifying commuters?
3. How accurately can commuters and non-commuters be identified using an ensemble model?
4. How many distinct behaviors are present among public transport users in Beijing?
5. How does feature selection and feature extraction compare to each other in the transportation domain?

3.1.1 Definition of terms

A commuter is a public transit user whose smart card data reveals repeatable patterns in time and space. Though commuters are usually associated with Monday to Friday 9:00am to 5:00pm schedules, in this work we extend the definition to any routine travel pattern. This flexibility allows us to include travelers with stable yet rare commuting schedules, such as night workers, weekend workers and evening workers.

A trip is a sequence of smart card transactions, including transfers, performed by the same user to travel from an origin to a destination. A trip is also represented as a record in the data, as it will be further explained in Section 4.1

A transfer is a change in transportation mode, or a change in vehicles whenever a smart card has to be checked within the same transportation mode. Transportation modes include Bus, Subway, and Bike.

We make the assumption that smart card IDs and users have a one to one relationship, meaning each user has exactly one card and each card is used by exactly one user. As discussed with domain expert Quian Tu, although some people may own more than one card, this is a minority. Thus, the assumption holds for the majority of travelers.

3.2 Scope and structure

This project is divided three main stages:

PART I: Prepare and preprocess the data using Big Data techniques In this part we focus on research question 1. Techniques for cleaning, knowledge extraction, categorization, patching and standardization are used and tailored to the data. From this, we build an appropriate 3 dimensional representation for each user's traveling behavior. This part corresponds to Section 5.

PART II: Classify commuters versus non-commuters by using an ensemble model In this part we focus on research questions 2 and 3. First, we perform feature selection in order to identify the most informative features and disregard redundant information. An extensive analysis of spatiotemporal properties is be done, combining transportation domain knowledge and statistical tools. Later, we create a classifier using ensemble models and discuss its performance. This part corresponds to Section 6

PART III: Users clustering according to patterns in their travel behaviors. In this part we focus on research question 4. First, we do feature extraction with the goal of reducing the dimensionality of the data. This is done via a convolutional autoencoder. Finally, we cluster the low dimensional representation using k-means clustering and do cluster analysis to understand the underlying pattern of each cluster. This part corresponds to Section 7

Figure 3 displays a flowchart for the stages and their connection.

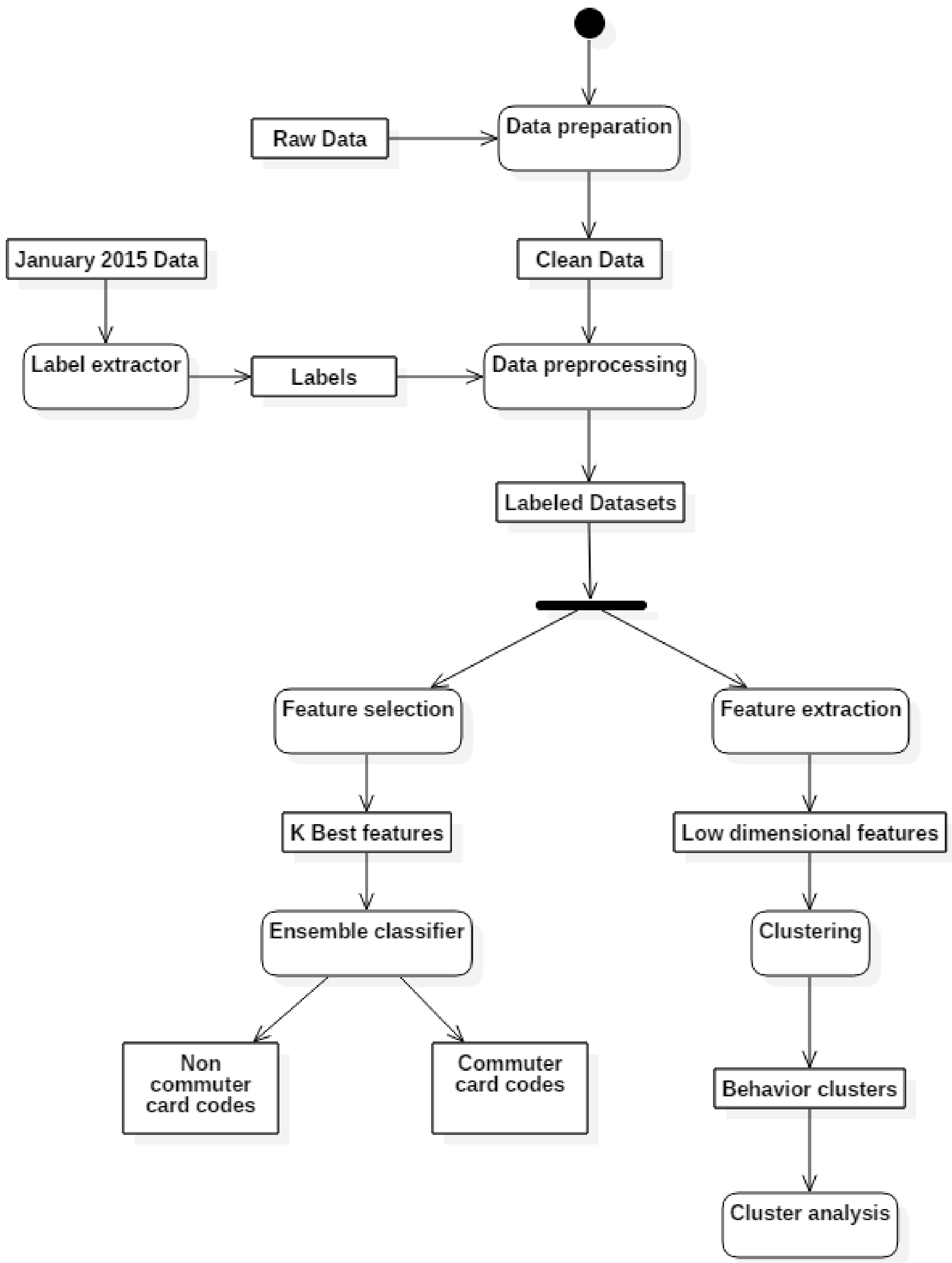


Figure 3: Project flow

4 Methodology

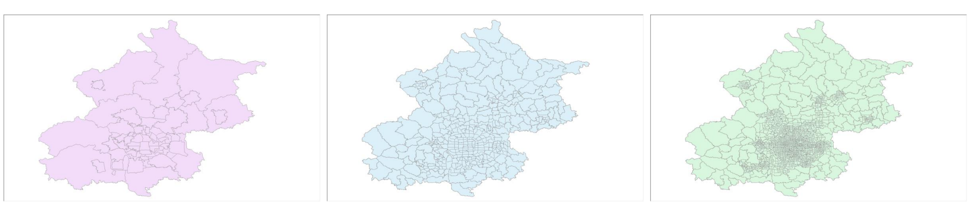
4.1 The data

Every record in the data represents a trip performed by a specific smart card. As such, it contains the following data fields:

- Data date: Year, month and day that the trip was made
- Card code: Card identification number
- Path link: Mode of transportation. B stands for bus, R for subway, Y for bicycle. Transfers between modes are shown by a dash. [1]
- Travel time: Time spent in vehicles, measured in milliseconds
- Travel distance: Distance traveled, measured in meters as performed by route.
- Transfer number: Number of changes in travel mode during the trip.
- Transfer total time: Total time spent in transfer, measured in milliseconds
- Transfer average time: Time spent in transfer, divided by number of transfers. Measured in milliseconds
- Start/End time: Time stamp of when the trip started/ended. Date and time up to milliseconds precision
- On/Off small traffic area: Integer ranging from 1 to 1911
- On/Off middle traffic area: Integer ranging from 1 to 389
- On/Off big traffic area: Integer ranging from 1 to 60
- On/Off ring road: Integer ranging from 1 to 6
- On/Off area: Integer ranging from 1 to 18
- ID: record identification number created by joining the following: hour of the beginning of the trip | time stamp of beginning of the trip | card code performing the trip
- Transfer detail: Mode of transportation, as well as line/route number and stations for boarding and alighting. More detail provided in Section 5.2.2

Full privacy of card users is ensured, as there is no personal data linking card codes to specific individuals.

The traffic zones (small, middle and big areas) are divided by the Beijing Municipal Institute of City Planning and Design (BICP). They are specific in different degrees, as shown in Figure 4. In general, the division principles correspond to the geopolitical environment and administrative planning, for example roads, villages and others. The 6 ring road and 18 areas districts are divided by the Beijing Municipal Government. The division is unique in Beijing. The 18 districts and counties are shown in Figure 5. According to domain expert PhD. Liang Quan, these divisions are sufficiently informative for traffic analysis [21].



(a) Big zones (b) Middle zones (c) Small zones

Figure 4: Traffic zone division

Every day, more than 13 million records are collected, with approximately 5 million corresponding to subway trips, 8 million corresponding to bus trips and 100,000 corresponding to bicycle trips.

[1] Example: B-B represents a Bus to Bus transfer.

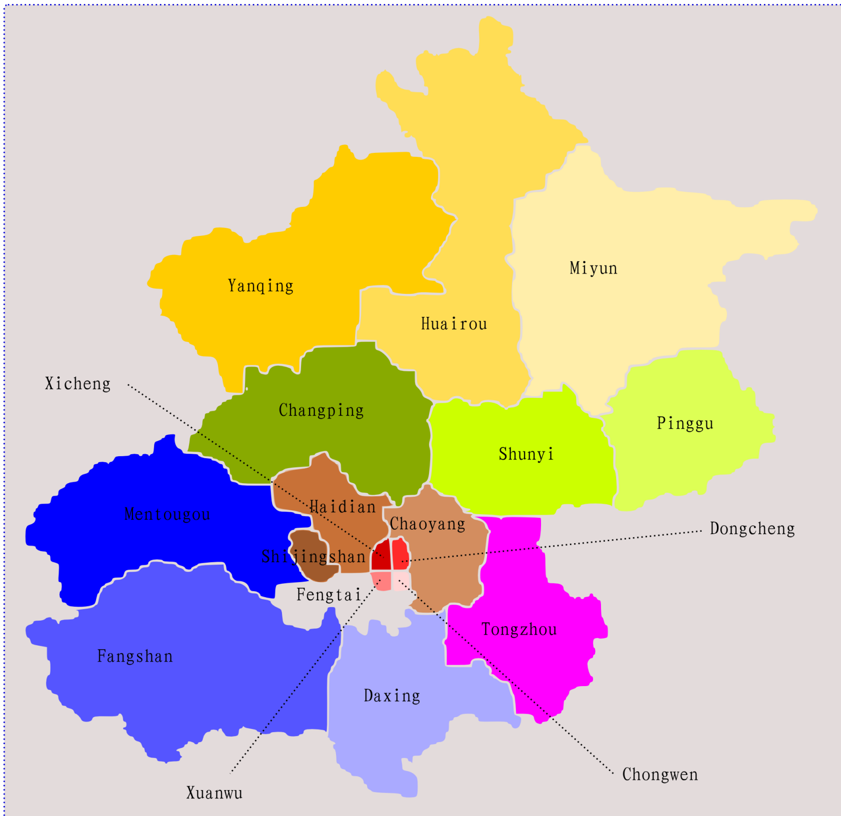


Figure 5: Beijing's Districts and its Counties

4.1.1 Special considerations

The previous description corresponds to the data as delivered by the Beijing Transportation Research Centre. As such, it is the result from processing the raw records at the collecting phase. Some special considerations concerning this processing are explained below:

Travel distance by bike: Since bicycles do not have predefined routes, the distance cannot be directly recorded. However, it is inferred by using the travel time and a static average speed for cyclists.

Subway transfer: Transfers between subways lines of the same operator cannot be tracked since a single check-in gives access to the traveler to all the subway network. In order to infer the transfer detail, the A* algorithm is used to calculate the most likely transfer sequence, given the boarding and alighting stations. Similarly to the bicycle missing information, the transfer time inside the subway system cannot be directly recorded. Using a static average walking speed and the known distance in transfer stations, the transfer time is calculated.

Transfer information: The path link and transfer number fields are extracted from the transfer detail field. Similarly, the transfer average time is calculated from the transfer total time and transfer number fields.

4.1.2 Labeled and unlabeled data

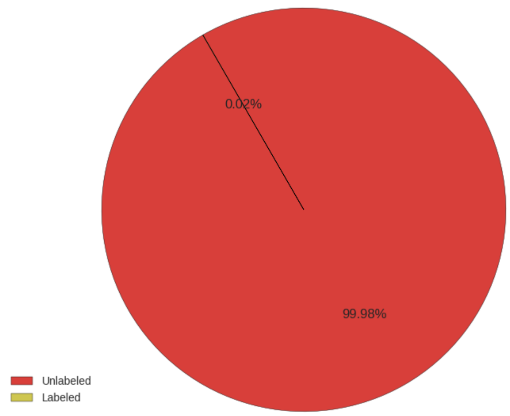


Figure 6: relation between labeled and unlabeled data.

Commuter classification: Classification is a supervised learning task, where every training data sample requires an associated label determining its true class. In case of commuter classification, this translates to having smart card codes associated with either a "commuter" or "non-commuter" label. Such data is expensive and limited since it can only be obtained by asking the users directly if they are commuters or not. Thus, in general, annotated data is not available, and labeling new records falls beyond the scope of this project.

As a solution for the above, we take advantage of the dataset used by Tu[23]. This dataset corresponds to trip records performed during a week in January 2015, and it contains labels for 978 smart cards, collected and validated via surveys. The original dataset distribution is composed by:

- 6439 records of 481 commuters
- 1628 records of 497 non-commuters

For this project, the Beijing Transportation Research Centre has provided us with one month worth of data, corresponding to January 2015. In order to construct an extended labeled dataset, we take these 978 labeled smart card IDs and search for their corresponding records in the one month sample. This dataset is used for Part I (Section 5) and Part II (Section 6) of this project.

Commuter clustering: In order to further cluster commuters, 100,000 smart card codes are sampled from data corresponding to November 2016. The month of November is chosen because it does not overlap with holidays and has a relatively stable weather thus diminishing the variance between bicycle and bus/subway traveler preferences. The year is chosen to reflect a more recent characterization of travelers.

4.2 Spatio-temporal representation

In this work we propose a 3 dimensional data representation to contain the monthly travel information of a user. This is shown in Figure 7.

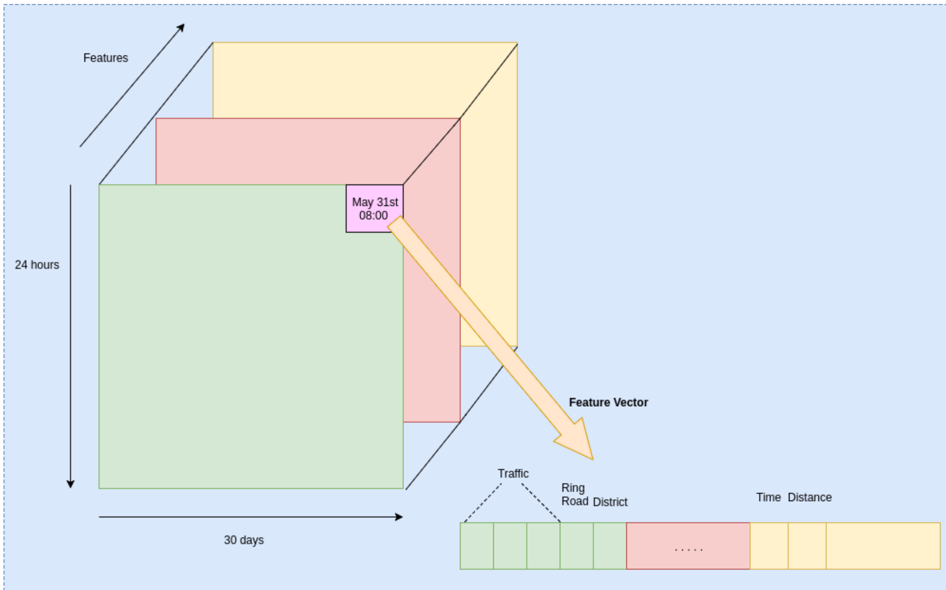


Figure 7: Spatio-temporal data structure.

Inspired by [10], the x-y plane constructs a temporal structure between days of the month and hours of the day. The crucial advantage of this structure lies in its local properties. Similar to the case of image processing, in this representation a temporal pixel is simultaneously influenced by what happened in the previous/following hours (y axis), and on the previous/following days (x axis).

As for the z plane, each layer contains a trip feature. In Figure 7 boarding spatial features are portrayed in green, alighting spatial are portrayed in red, and other types of features (such as travel time, travel distance, transfer number, transfer total time, etc) are portrayed in yellow.

Therefore, each temporal pixel may contain a trip feature vector, which expands several layers deep. Considering that even regular public transport users do not perform more than 6 trips a day as shown by the number of trips distribution in Figure 9, the proposed representation is sparse, since only a few time pixels are populated with trips.

4.3 Dimensionality reduction

Considering there are about 26 attributes in a trip, we have $24 \times 30 \times 26 = 18,790$ temporal pixels per user. Given the high dimensionality and the sparsity of the structure, we will perform dimensionality reduction.

4.3.1 Feature selection

techniques for choosing best k

Statistical such as correlation, chi squared, anova Machine learning such as trees Domain knowledge

4.3.2 Feature extraction

Mapping between high dimensional and low dimensional through autoencoders.

Taking advantage of the local properties of the proposed structure, we can apply convolutional filters [2] to reduce the dimensionality to a more manageable number. The end result will be used as features for clustering commuters in Part III (Section 7) of this project.

[2] CNN chosen because of local properties, is PCA also local?

4.4 Pattern recognition

4.4.1 Ensemble models

Ensemble models benefit from combining non-correlated prediction methods. Weak classifiers might correct each other in specific hard cases. Ensemble models are chosen for this project because of its robustness and modularity. Starting from a few simple classifiers, assembled via aggregation methods, the model can grow larger or more complex as needed.

Supervised learning As proven by Tu [23], weak classifiers like an SVM prove to be sufficient to identify commuters. This hints to extend the ensemble model with other similar weak classifiers like decision trees, Bayesian classifiers or multilayer perceptron. Bagging will be used to ensemble their predictions.

4.4.2 Clustering

K means

5 Data preparation and preprocessing

5.1 Cleaning

As first step for preprocessing the data, we eliminate faulty records. The different filters are:

1. Eliminate records with missing data: 10.93% records eliminated
2. Eliminate records with missing travel details: 1.58% records eliminated
3. Eliminate records with travel time <= 0: <0.01% records eliminated
4. Eliminate records with travel distance <= 0: 9.82% records eliminated

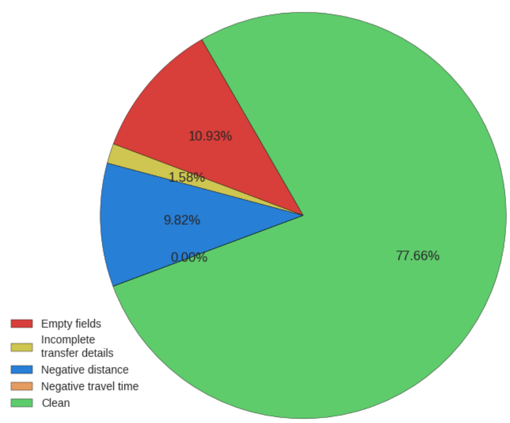


Figure 8: Reasons for eliminating records.

The first four filters aim to eliminate records with missing fields, which already reduces the dataset to 77.66% of its original size. [3]

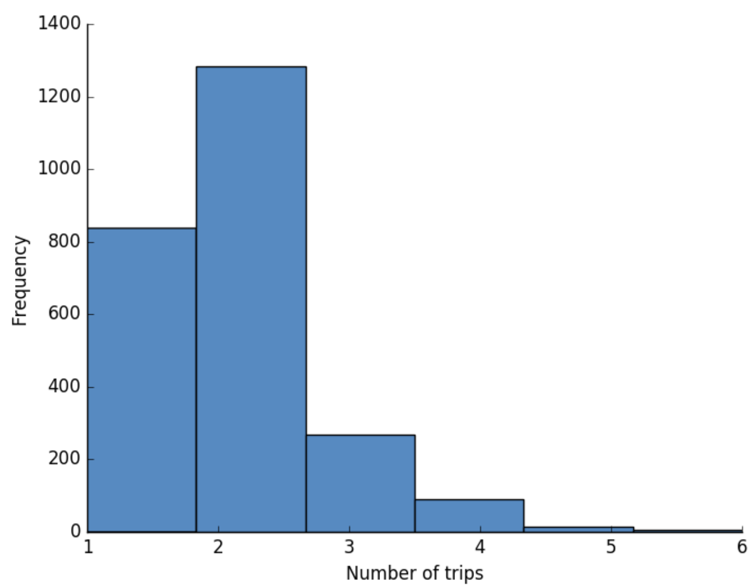


Figure 9: Number of trips distribution. 2500 record sample.

[3] explain figure

Figure 9 shows the distribution of number of days in a single day. Its shows that most people perform two trips per day. Fixing the minimal number of trips to 60, which is equivalent to an average of two trips per day, the final dataset contains 51.76% records available for usage.

5.2 Extraction

5.2.1 Time bins

Regardless of its criticism, using hourly time bins is standard practice in the field and has shown sufficient to examine temporal data [10] [11] [15]. Therefore, in this project we follow the same technique and extract only the hour of the start and end of each trip.

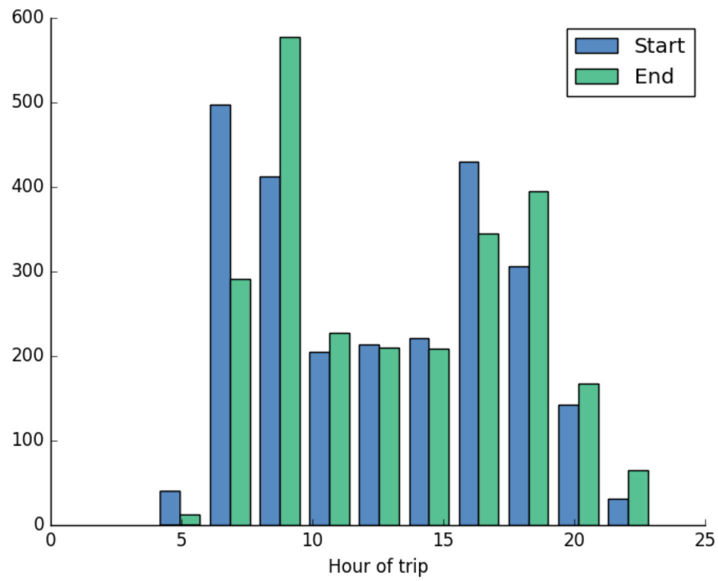


Figure 10: Distribution of start/end hours for trips. 2500 records sample.

From Figure 10, we note that our data follows the expected distribution for the domain, showing clear morning and evening peaks. Furthermore, we note that boarding and alighting patterns during the peaks hours are shifted by one hour. This is explained by the previous finding, that the mean travel time is almost one hour.

5.2.2 Trip parsing

The trip details obtained from the records are given in Chinese, with descriptors containing a combination of numbers and text. In order to extract boarding/alighting route features, the descriptors must be parsed.

We parse the trip details using a combination of two techniques: regular expressions and tokenization.

Regular expression Since a trip may include transfers, we define a trip to be composed of one or many rides. Each ride is carried out in a single travel mode.

In order to obtain the elements of each ride we look at the pattern per travel mode.

$$\begin{aligned} BIKE &= (bike.STOP - STOP) \\ SUBWAY &= (subway.LINE : STOP - LINE : STOP) \\ BUS &= (bus.ROUTE(DIRECTION - DIRECTION) : STOP \\ &\quad - ROUTE(DIRECTION - DIRECTION) : STOP) \end{aligned}$$

where the upper-case text corresponds to placeholders for ride elements, the lower-case text corresponds to the English translation of the descriptor in Chinese, and the punctuation (parentheses, dots, colons and dashes) correspond to separators between ride elements.

Unifying the mode-specific patterns, we describe a ride and a trip using regular expressions:

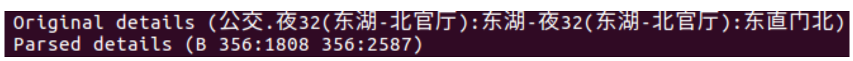
$$RIDE = (MODE.[LINE/ROUTE:]?STOP - [LINE/ROUTE:]?STOP)$$
$$TRIP = RIDE[->RIDE]?$$

where elements surrounded by squared brackets and followed by a question mark (e.g. $[ELEMENT]?$) correspond to optional elements. We note that when parsing bus details, we disregard the route direction. This decision is motivated to fit both subway lines and bus routes to a single pattern, noting that the direction of the route does not affect the path of the route itself.

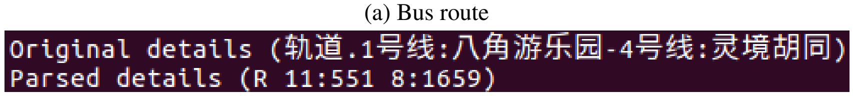
Tokenization Once the elements of a trip are extracted, they must be substituted with numerical IDs. These IDs are not available from the Beijing Institute of Transportation, thus three different vocabularies are created for subway lines, bus routes and combined stops correspondingly.

Usually, bus routes are identified by a number. However, in Beijing a single bus route number can be associated to different paths. Such is the case of night, express and special cases of a bus route, which follow different paths even if they are described with the same number. For this reason we create a vocabulary with all unique parsed routes according to their full description and not only their number.

Examples of cleaned routes are shown in Figure 11



(a) Bus route



(b) Subway route

(c) Bike route

Figure 11: Examples for parsed and tokenized trip details.

5.3 Data patching

We note that the number of transfers and the path link fields of some records do not correspond to the information in their trip details. According to domain expert PhD. Tu Qiang, this must be recalculated [20]. Figure 12 shows the distribution of the number of transfers per trip before and after patching. [4]

[4] piechart might be better?

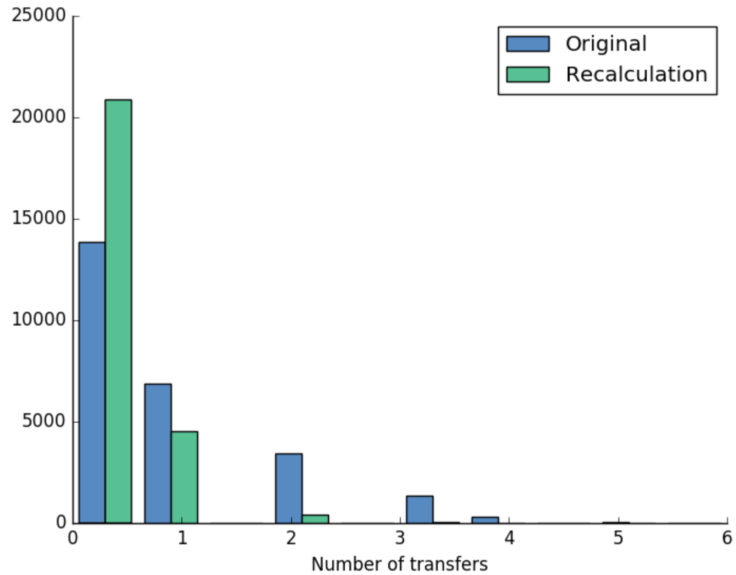


Figure 12: Transfer number distribution before and after recalculation.

Our distribution shows that most trips are performed without transfers, which is consistent with other studies findings [1].

5.4 Standardization

In data mining, it is a standard practice to perform whitening. This technique eliminates correlations between features, which is desirable in most cases. However, the domain of Metropolitan Transportation some of these correlations are highly important, and should not be discarded. This is the case of total travel time and distance, as shown in Figure 13. For this reason, we choose to only standardize the features and keep the correlations.

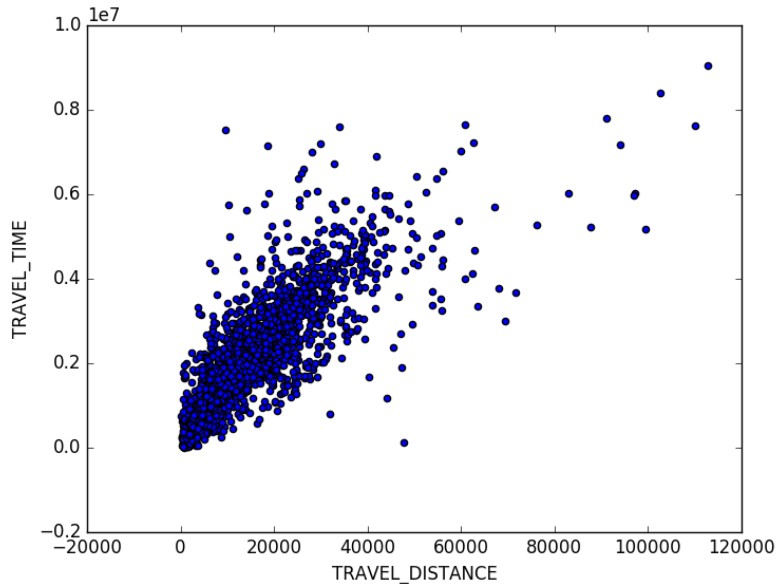
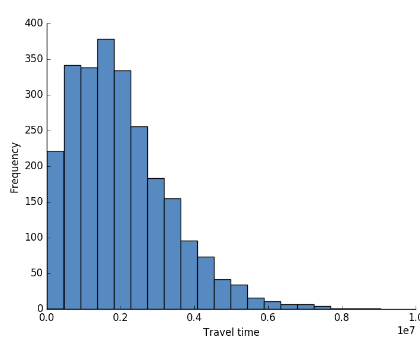
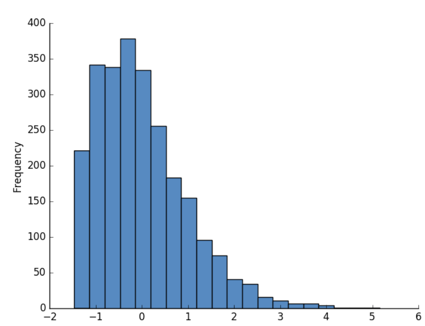


Figure 13: Travel distance vs travel time. 2500 record sample.

Travel time, travel distance, total transfer time and average transfer time where standardized by subtracting the mean of each distribution and forcing a unit standard deviation.

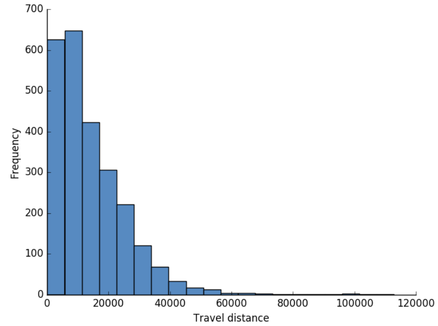


(a) Original

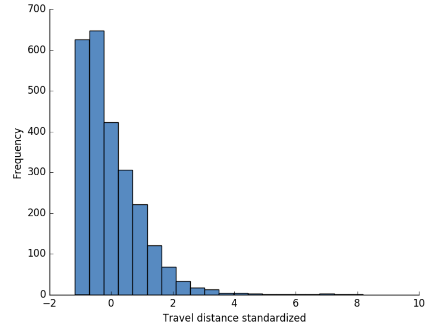


(b) Standarized

Figure 14: Time distribution before and after preprocessing. 2500 records sample.



(a) Original

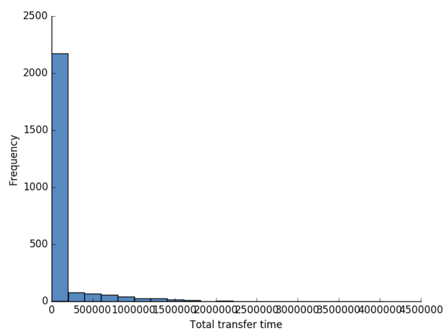


(b) Standarized

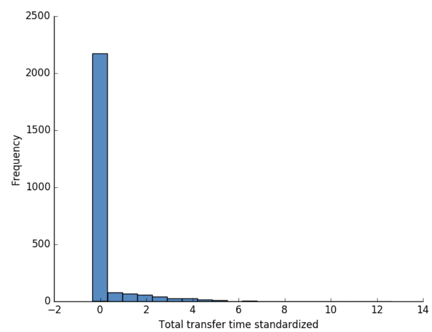
Figure 15: Distance distribution before and after preprocessing. 2500 records sample.

Figure 14 and 15 show travel time and distance follow a truncated Gaussian distribution $\mathcal{N}(\mu = 1, \sigma^2 = 1)$. Since the nature of the data prevents negative values (time and distance must be positive), the original distribution is truncated at 0. Standardization maintains the shape of the distribution, but shifts and contracts it to be closer to zero values.

The mean travel time is 55 minutes, and the mean travel distance is 10 kilometers.

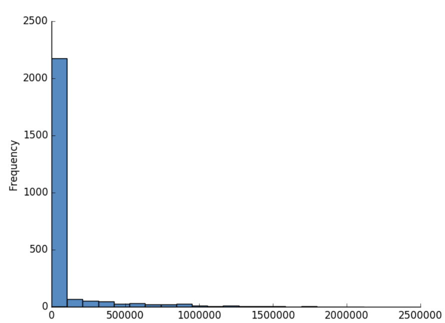


(a) Original

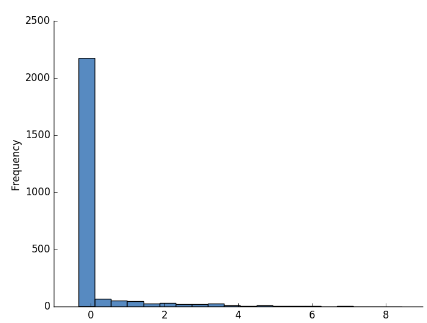


(b) Standarized

Figure 16: Total transfer time distribution before and after preprocessing. 2500 records sample.



(a) Original



(b) Standarized

Figure 17: Average transfer time distribution before and after preprocessing. 2500 records sample.

Figure 16 and 17 show that transfer times, both total and average, follow distributions with very long tails. As mentioned before, most trips are performed without transfers, which explains that most of the trips have transfer times equal to zero. [5]

5.5 Attributes

The spatial features to be included in the representation are:

1. Small traffic area
2. Middle traffic area
3. Big traffic area
4. Ring road
5. District
6. Mode
7. Line/Route
8. Stop

These are repeated for boarding and alighting information.

The general trip features to be included in the representation are:

1. Travel time
2. Travel distance
3. Transfer number
4. Transfer total time
5. Transfer average time
6. Start hour
7. End hour
8. Number of trips

The attributes are of two types: numerical and categorical. Usually categorical should be done by one hot encoding. In this case it is not scalable.

[5] do not include transfer 0. focus on tail behavior

5.6 User cubes

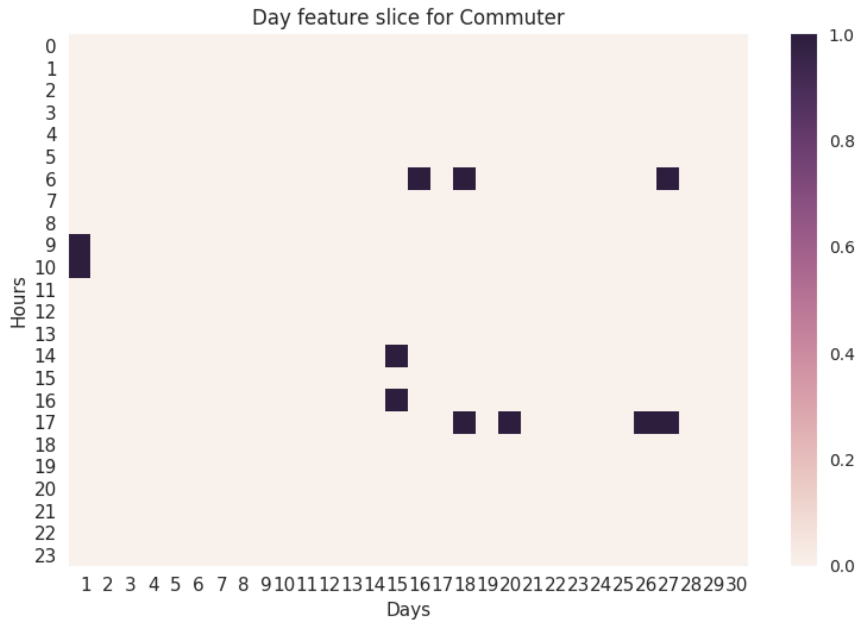


Figure 18: Sample Day slice for random Commuter.

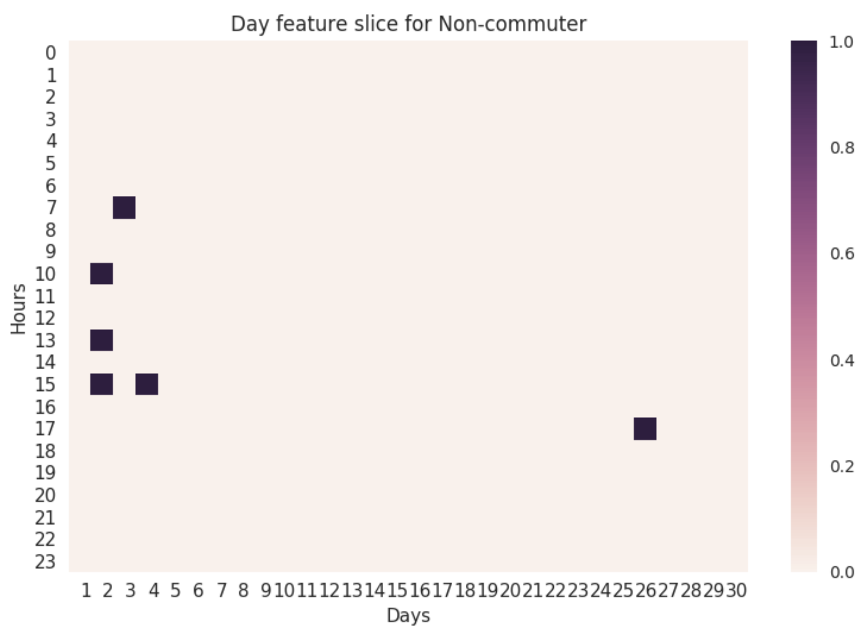


Figure 19: Sample Day slice for random Non-commuter.

6 Commuters identification

6.1 Attributes correlation

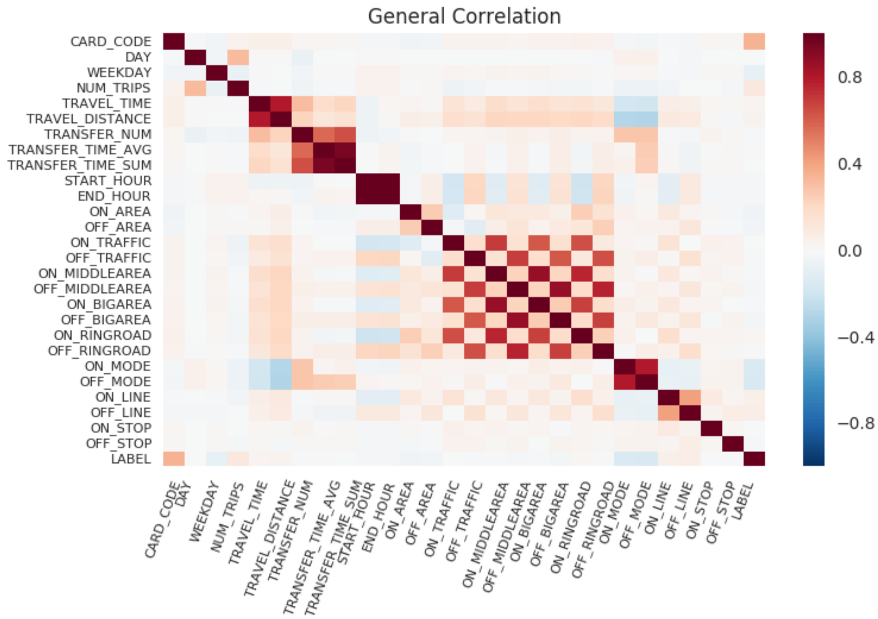


Figure 20: Attributes correlation to each other and to label.

General

Temporal

Spatial: Hierarchy in geographical areas causes correlation

6.2 Feature selection

Statistical correlation to label, f value of anova. chi 2 discarded because it focuses on times only machine learning through trees domain knowledge

Aggregation system. Each method's scores are normalized to add up to 1. Each contribute with equal weight.

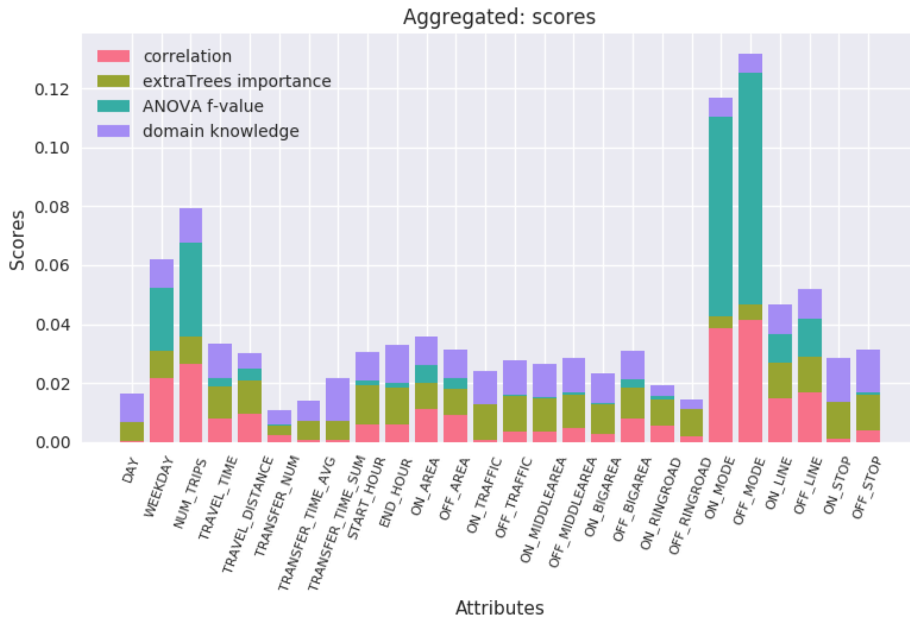


Figure 21: Attributes scores.

Choose middle area and mode.

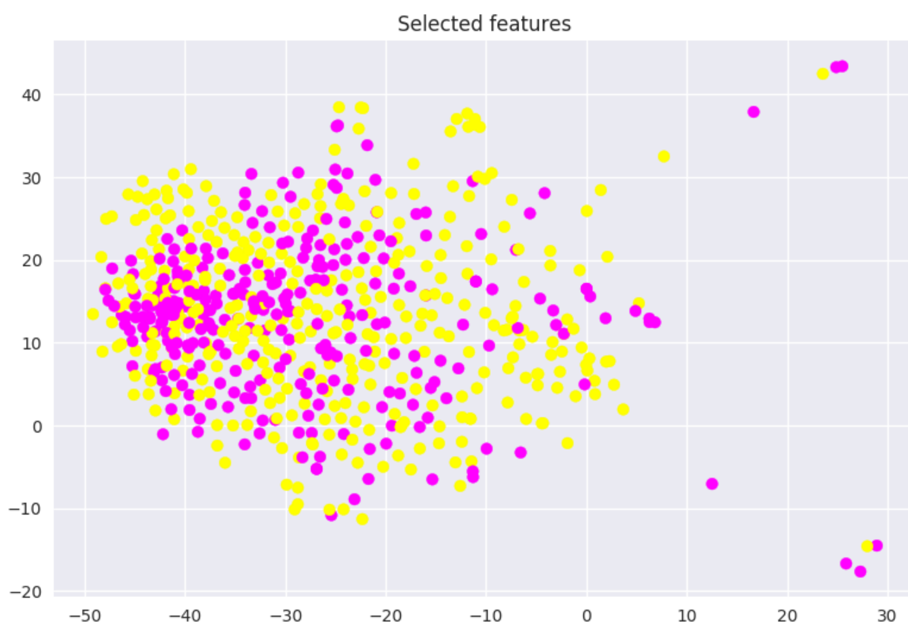


Figure 22: Visualization of samples and their labels after feature selection.

6.3 Model

Take slices of cubes and flatten.

We compare a single SVM, ensembled random forests.

As suggested by Tu [23] results, the data is almost linearly separable thus simple classifiers such as decision trees may suffice.

6.4 Experiments

SVM

Accuracy: 94.66%

Figure 23: SVM confusion matrix.

Random Forest

Accuracy: 97.54%

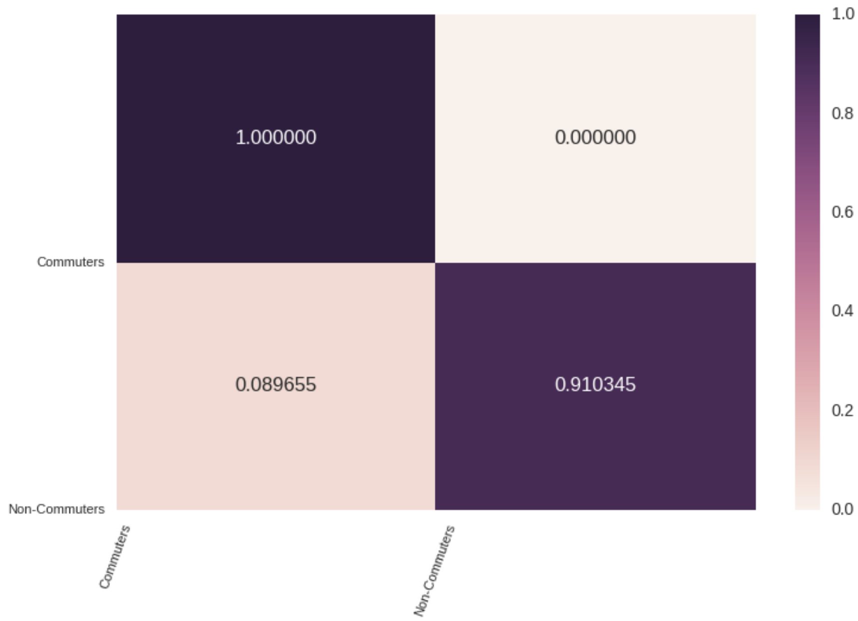


Figure 24: Random forest confusion matrix.

AdaBoost

Accuracy: 98.24%

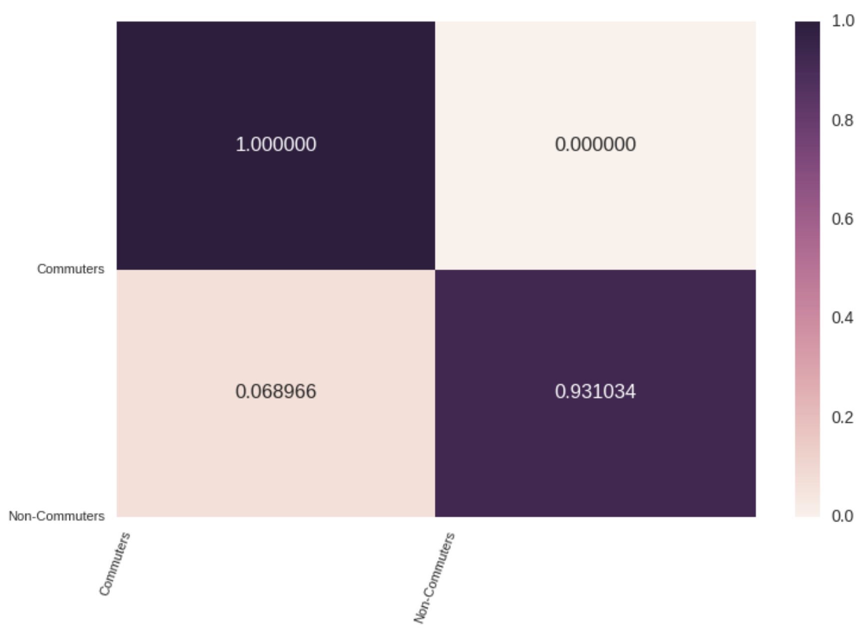


Figure 25: Random forest confusion matrix.

6.5 Discussion

Random Forests deal well with categorical information. Adaboost deals well with missing information.

Overfitting

7 Commuters clustering

7.1 Feature extraction

The features to be fed into the clustering algorithms are obtained by dimensionality reduction. As mentioned before, convolutional filters can exploit local information.

7.1.1 Convolutional filters

In order to examine weekly patterns, our the first layer of convolutional filters will have an x dimension of 7 with a stride of 7. Assuming that only the hour previous and after a trip affects the trip, the filter will have a y dimension of 3 with a stride of 1. We do not perform padding, since this would have significant implications in the travel behavior of the user.

Following this formula:

$$output = \frac{input - filter}{stride} + 1$$

We find that the output size after the first convolution is 4×22. Considering 15 features, this reduces the dimensionality to $4 \times 22 \times 15 = 1320$

Goal is to have a $4 \times 3 \times 3 = 36$ structure

[6]

7.1.2 Autoencoder

A neural network is constructed to apply the convolutional filter. After the convolution layer, ReLU is applied as activation function.

TSNE results

7.2 Clustering

Tuning

Evaluation

7.3 Cluster analysis

7.4 Discussion

[6] what about depth? can I have a filter that is 8 units deep, with 8 stride (features are divided as 8 spatial boarding, 8 spatial alighting and 8 transfer related)? How could this be applied?

8 Conclusion and future work

Faulty data. Missing data from 9 to 16. Gap in between.

Zero values such as end hour, zero transfers and others. Same effect as "empty" bin.

Overlapping trips. So far one trip per bin/hour. Possible solutions are to concatenate.

Standardization by day.

Selection over trip -> label.

References

[1] Ashish Bhaskar, Edward Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3):1537–1548, 2015.

[2] Philip T Blythe. Improving public transport ticketing through smart cards. In *Proceedings of the Institution of Civil Engineers, Municipal Engineer*, volume 157, pages 47–54. Citeseer, 2004.

[3] Beijing Transportation Research Center. 2016 annual report on traffic development in beijing. `http://www.bjtrc.org.cn/InfoCenter/NewsAttach/2016%E5%B9%B4%E5%8C%97%E4%BA%AC%E4%BA%A4%E9%80%9A%E5%8F%91%E5%B1%95%E5%B9%B4%E6%8A%A5_20161202124122244.pdf`, 2016. Acessed on 21 April, 2017.

[4] Mo Lim Chan. *Tactical implementation model for the smart card payment system for metro operation*. PhD thesis, City University of Hong Kong, 2010.

[5] Patrick YK Chau and Simpson Poon. Octopus: an e-cash payment system success story. *Communications of the ACM*, 46(9):129–133, 2003.

[6] Gerhard de Koning Gans. *Analysis of the MIFARE Classic used in the OV-chipkaart project*. PhD thesis, Master's thesis, Radboud University Nijmegen, 2008.

[7] Kamran Ghasedi Dizaji, Amirhossein Herandi, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *arXiv preprint arXiv:1704.06327*, 2017.

[8] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[9] US EIA. Energy information administration (2016), international energy outlook 2016, with projections to 2040. Technical report, DOE/EIA-0484, 2016.

[10] Gabriel Goulet Langlois, Haris N Koutsopoulos, and Jinhua Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.

[11] Xiaolei Ma, Congcong Liu, Huimin Wen, Yunpeng Wang, and Yao-Jan Wu. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58:135–145, 2017.

[12] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[13] Zidan Mao, Dick Ettema, and Martin Dijst. Commuting trip satisfaction in beijing: Exploring the influence of multimodal behavior and modal flexibility. *Transportation Research Part A: Policy and Practice*, 94:592–603, 2016.

[14] David Marr and A Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, 1(2), 1982.

[15] Catherine Morency, Martin Trepanier, and Bruno Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[16] D Kanishka Nithin and P Bagavathi Sivakumar. Generic feature learning in computer vision. *Procedia Computer Science*, 58:202–209, 2015.

[17] OECD. Passenger transport (indicator). `10.1787/463da4d1-en`, 2017. Acessed on 10 April, 2017.

[18] Beijing Municipal Commitee of Communications and Beijing Transportation Reserch Center. Research on beijing's public transportation commuting transit network. `https://www.esmap.org/sites/esmap.org/files/10282009102930_Beijing_Transport_finalReport.pdf`, 2009. Acessed on 21 April, 2017.

[19] Meisy Andrea Ortega-Tong. *Classification of London's public transport users using smart card data*. PhD thesis, Massachusetts Institute of Technology, 2013.

[20] Tu Qiang. personal communication.

[21] Liang Quan. personal communication.

[22] World Population Review. Beijing population. http://worldpopulationreview.com/world-cities/beijing-population/, 2016. Acessed on 21 April, 2017.

[23] Qiang Tu, Jian-cheng Weng, Rong-Liang Yuan, and Peng-fei Lin. Impact analysis of public transport fare adjustment. *Traffic Engineering & Control*, 57(2), 2016.

[24] UITP. World metro figures, statistics brief. http://www.uitp.org/sites/default/files/cck-focus-papers-files/UITP-Statistic%20Brief-Metro-A4-WEB_0.pdf, 2015. Acessed on 21 April, 2017.

[25] Vukan R Vuchic. Urban public transportation systems and technology. 1900.

[26] Kunyan Wang, Qiong Luo, and Xueying Zang. Studies on ecological environmental carrying capacity in beijing, tianjin, and hebei region. In *Report on Development of Beijing, Tianjin, and Hebei Province (2013)*, pages 119–137. Springer, 2015.

[27] Jiancheng Weng, Yueyue Wang, Jianling Huang, and Ledian Zhang. Bus operation monitoring oriented public transit travel index system and calculation models. *Advances in Mechanical Engineering*, 2013.

[28] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487, 2016.

[29] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2016.

[30] Hefeng Zhang, Shuxiao Wang, Jiming Hao, Xinming Wang, Shulan Wang, Fahe Chai, and Mei Li. Air pollution and control action in beijing. *Journal of Cleaner Production*, 112:1519–1527, 2016.

[31] Shaojun Zhang, Ye Wu, Huan Liu, Ruikun Huang, Liuhanzi Yang, Zhenhua Li, Lixin Fu, and Jiming Hao. Real-world fuel consumption and co 2 emissions of urban public buses in beijing. *Applied Energy*, 113:1645–1655, 2014.

[32] Jiangping Zhou, Enda Murphy, and Ying Long. Commuting efficiency in the beijing metropolitan area: an exploration combining smartcard and travel survey data. *Journal of Transport Geography*, 41:175–183, 2014.