NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Exploring city mobility through travel diaries: a non-negative tensor factorization approach

Nevena Cukrov

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Exploring city mobility through travel diaries: a non-negative tensor factorization approach

by

Nevena Cukrov

Master Thesis presented as a partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics.

Supervised by

Professor Fernando José Ferreira Lucas Bação

Month, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 14.07.2024.]

ABSTRACT

In the EU 64% of everyday people's transport is made by private vehicles (Palm, 2022.), but that leads to today's CO2 emissions of private road vehicles being responsible for 74% of all emissions related to transport (IEA. 2020). That's why there should be an effort to enhance the usage of public transportation, which was agreed upon by all the EU members who signed the Green New Deal (European Environment Agency, 2021). In this research, the status of the public transit network will be explored, and the pain points will be exploited to recognize the necessary changes required to facilitate this shift. This will be done using a non-negative tensor factorization approach which will help to better understand complex and multidimensional travel diaries.

KEYWORDS

Travel diary; Tensor factorization; Mobility patterns; Non-negative Tucker; Non-negative PARFAC

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In contemporary urban planning, one of the most significant challenges is designing and sustaining cities with a strong focus on sustainability and ecological considerations. This has led to the emergence of a crucial strategy: leveraging the big amounts of available data for more informed decision-making, in regards to urban planning. (Engin et al., 2020). Creating better public transportation systems is a pivotal aspect of this strategy. It directly contributes to reducing carbon emissions generated by private vehicles. The shift towards more sustainable modes of transportation is emphasized in the Paris Agreement, The agreement underscores the necessity of improving public transit to meet global climate goals. (Agencia Europea de Medio Ambiente., n.d.)

Road vehicles are responsible for a substantial portion of CO2 emissions connected to transportation. The International Energy Agency (IEA) reports that road vehicles contribute 74% of these emissions (IEA, 2020), and as urban areas continue to grow and evolve, the complexity of transportation networks and the diversity of urban behaviors present new challenges for planners and policymakers. The optimization of public transportation schedules based on actual usage patterns offers a promising solution to these challenges. The availability of digital traces, such as mobile phone data, GPS data, smart card data, and survey data, provides a new perspective for addressing these issues, provided there is effective collaboration between data science and transportation planning.

Understanding mobility patterns, characterized by their temporal and spatial features, is essential for identifying high-demand areas and times, facilitating more efficient resource allocation and service planning (Curado et al., 2021)). Insights gained from mobility pattern analysis can inform decision-makers and urban planners about necessary adjustments or improvements in public transportation networks (Fabbiani et al., 2018). Much of the existing research focuses on several key goals: identifying where and when the demand for transport is not satisfied by existing services, assessing the need to revise or extend transport services, and evaluating the benefits of shifting travel demand to public transport to reduce CO2 emissions and operational costs (Hadjidimitriou et al., 2021).

In this research, we will closely examine the topic of public transportation with a focus on optimizing public transportation schedules based on actual usage patterns. Digital traces from various data sources, such as mobile phone data, GPS data, smart card data, and traditional surveys, offer new perspectives for solving these challenges, provided there is effective collaboration between data science and transportation planning.

Matrix and tensor factorization techniques, particularly Non-Negative Tensor Factorization (NTF), have proven successful in representing original data through lower-dimensional approximations. Unlike traditional matrix factorization methods, NTF can handle multi-dimensional data, making it an effective solution for pattern recognition and urban planning

problems (Kim & Park, 2012). Ensuring non-negativity in the models improves interpretability, which is crucial for deriving meaningful insights from complex datasets.

NTF is a powerful tool in various fields, including biomedical informatics, signal processing, computer vision, and data mining, due to its ability to extract hidden patterns from high-dimensional data. For example, Welling and Weber (2001) used NTF to identify patterns in muscle activation signals, providing insights into muscle functions and aiding in diagnosing neuromuscular disorders. Similarly, Mørup et al. (2008) applied sparse NTF to brain imaging data, revealing significant neural activity patterns and enhancing the understanding of brain functions. This concept is also used with geo-spatial data to find patterns of movements, what will be the focus of this study.

This thesis research is based on travel diary data collected by the New York City Department of Transportation (NYCDOT). The Citywide Mobility Survey (CMS) collects data on travel behavior, preferences, and attitudes of New York City residents. This data is rich with information on how residents travel, including their reasons and preferred modes of transportation.

The primary objective of this research is to identify movement patterns in citizens' daily trips, with a focus on trips taken using public transportation. Traditional clustering algorithms often struggle with hyperparameter settings and noise in real-world data, failing to capture intricate spatial correlations. The proposed method, using NTF, aims to extract relevant urban patterns, helping urban planners and policymakers in decision-making processes. Two main algorithms, Non-negative PARAFAC and Non-negative Tucker, were evaluated for their effectiveness in pattern extraction and interpretability.

This paper will be split into 5 sections. In the next one, Literature Review, we will explain theoretical importance of this research, talk about the methods and algorithms that will be used, and showcase previous work with similar goals to prove the benefits of this research. Then in the Methodology section, we will go through the flow of research, from getting the data to results. In the fourth section, the results will be shown and examined. The finale sections will be Conclusion and Limitation & Future work.

2. LITERATURE REVIEW

In contemporary urban planning, a significant challenge is the struggle to design and sustain cities with a strong focus on sustainability and ecological considerations. This challenge has led to the emergence of a crucial strategy: leveraging the vast amounts of available data for more informed decision-making. (Engin et al., 2020)

In this paper, we will look closely at the topic of public transportation with a focus on optimizing public transportation schedules based on actual usage patterns. Cities' increasing complexity and changing urban behaviors pose new challenges for transportation planning. The availability of digital traces, like mobile phone data, GPS data, Smart Card data, and survey data, offers a new perspective for solving these challenges, provided there's effective collaboration between data science and transportation planning.

Collaboration and smart public transport network design offer several ecological benefits. They reduce traffic overcrowding, as already mentioned. Efficient public transportation systems can significantly reduce the number of private vehicles on the road, alleviating traffic congestion and lowering overall emissions. Another benefit would be better usage of energy. Public transportation systems, especially electric buses and trains are generally more energy-efficient than individual car use. This efficiency translates to lower fuel consumption and reduced greenhouse gas emissions. Promoting active transport could help citizens live healthier lives. Integrating public transport with pedestrian and cycling infrastructure encourages active modes of transport, which are not only eco-friendly but also promote public health. Smart public transportation network, also means better land use. Well-designed public transport networks can influence urban development patterns, promoting higher density and mixed-use developments. This can reduce urban sprawl, preserve green spaces, and lower the environmental impact of urban expansion (Engin et al., 2020)

One way of exploring mobility data, that would be beneficial for urban network planers is extracting mobility patterns. Mobility patterns are characterized by their temporal and spatial features. (Curado et al., 2021) Studying mobility patterns enables the identification of high-demand areas and times, facilitating more efficient resource allocation and service planning. The insights gained from mobility pattern analysis can inform decision-makers and urban planners about necessary adjustments or improvements in the public transportation network. (Fabbiani et al., 2018)

Much of the research has been published on the topic, and most of it focuses on these goals: (1) to quickly understand where – between which traffic zones – and when – during which time ranges – the demand for transport is not satisfied by existing public transport services; (2) to understand whether there is the need to revise or extend an existing transport service (e.g., a bus route, a metro line); (3) to assess the maximum benefits of shifting travel demand to public transport, keeping in mind CO2 reduction and operational cost. (Hadjidimitriou et al., 2021)

2.1. MATRIX/TENSOR FACTORIZATION

Matrix Factorization, or in this case Tensor Factorization, is the method used to represent the original data by a lower-dimensional approximation obtained via matrix or tensor (multiway array) factorizations or decompositions. The notion of matrix/tensor factorizations arises in a wide range of important applications and each matrix/tensor factorization makes different assumptions regarding component (factor) matrices and their underlying structures. So, choosing the appropriate one is critical in each application domain. Approximate low-rank matrix and tensor factorizations fundamentally enhance the data and extract latent (hidden) components. (Zhou et al., 2014)

Nonnegative Matrix Factorization (NMF) involves approximating a nonnegative matrix by expressing it as the product of two nonnegative matrices. Typically, these resulting matrices are smaller than the original one, simplifying its handling and comprehension. It was first introduced as an idea, in 1994, by researcher Pentti Paatero, Unto Tapper. This method was introduced as a competitive method for principal component analysis (PCA) and factor analysis (FA). The newly introduced method relies on error estimates of the elements within the measured data matrix. It enforces strict non-negativity constraints on the factors, both for the basis and the coefficients utilized in approximating the matrix. (Paatero & Tappert, 1994)

The method was widely recognized a few years later, after Daniel D. Lee and H. Sebastian Seung, published their research on the topic in the journal in 1999. Since then, this method has been studied more closely for many topics some of which are already mentioned clustering and pattern recognition, but it is also used in text mining, classification of documents and emails, spectral data analysis, and face recognition. (Fan et al., 2014)

The literature describes Nonnegative Matrix Factorization (NMF) in the following manner: Consider a nonnegative matrix **Y** with dimensions I × T. The goal is to identify two nonnegative matrices, **A** (I × J) and **X** (J × T), whose product, **AX**, closely resembles **Y**. A matrix is deemed nonnegative when all its entries are 0 or positive. Typically, the value of J selected is significantly less than both I and T. It's important to acknowledge that finding **A** and **X** such that **AX** exactly equals **Y** is generally not possible. Therefore, NMF is considered an approximation method, sometimes referred to as Approximative Nonnegative Matrix Factorization or Nonnegative Matrix Approximation. This concept is often articulated as **Y = AX + E**, where **E** represents an error matrix of dimensions I × T, highlighting the approximation discrepancy. Consequently, **AX** offers a condensed representation of **Y**, possessing a rank no greater than **J**. (Fan et al., 2014)

Given a nonnegative matrix $Y \in R^{I \times T}$ and a positive integer J, find non-negative matrices $A \in R^{I \times J}$ and $X \in R^{J \times T}$ that minimizes the function.

$$f(A, X) = \frac{1}{2} \| Y - AX \|_F^2$$

The goal of NMF is to find matrices A and X that minimize the value of the function f(A,X), which is defined as half the Frobenius norm of the difference between Y and AX. The Frobenius norm

is a measure of matrix size that is the square root of the sum of the absolute squares of its elements.

$$J < \min\{I, T\}$$

While it's stated that the integer **J** should be less than the minimum of **I** and **T**, this isn't a strict requirement but is commonly the case in practical applications. By choosing **J** smaller than **I** and **T**, the matrices **A** and **X** provide a lower-dimensional representation of the original matrix **Y**.

For an I × T matrix M, $\| M \|_F$ is the Frobenius norm of M, defined as:

$$\| M \|_F = \sqrt{\sum_{i=1}^{I} \sum_{t=1}^{T} m_{i,t}^2}$$

where $m_{i,t}$ denotes the element of **M** with row index **i** and column index **t**. Therefore, f(A, X) is the square of the Euclidean distance between Y and AX with an additional factor. The problem of finding **A** and **X** is convex when considering **A** and **X** separately, meaning that local minima are also global minima for each matrix individually. However, the problem is not convex when considering both **A** and **X** together, which can make finding the global minimum more challenging.

This approach showcases using Euclidian distance to find the distance between **Y** and **AX**. But it is also possible to solve the same problem using different algorithms such as Kullback-Leibler divergence, Csiszar's divergences, or Alpha- and Beta-divergences. Using different metrics results in distinct NMF algorithms, or at the very least, alters the update procedures within those algorithms. (Fan et al., 2014)

Matrices are considered two-dimensional tensors. In certain situations, such as analysis involving multi-dimensional data, the input consists of tensors that are of third order or higher. Consequently, there is a need to extend the principles of Nonnegative Matrix Factorization to accommodate the factorization of higher-order tensors, leading to the concept of Nonnegative Tensor Factorization. (Kim & Park, 2012) Instead of factoring a matrix into two nonnegative matrices, NTF factors a tensor into a set of nonnegative component matrices.

There are several tensor decomposition methods. The two most used are PARAFAC, also known as CANDECOMP, and TUCKER. Both methods are multi-linear decomposition techniques that break down the array into sets of factors and weights, so they can represent the original data, but in a compact structure. PARAFAC uses factorization to decompose a tensor into three two-dimensional components or matrices. And, TUCKER uses principal component analysis, resulting in different decomposition structures. It produces three two-dimensional components along with an additional core matrix that links the components. This core matrix makes interpreting data using the TUCKER model more complex (due to the

increased number of parameters) compared to PARAFAC. (Bro, 1997), (Rošť Áková et al., 2020).

2.1.1. The non-negative PARAFAC model

Given a tensor **Y** of order N, with nonnegative entries and dimensions $I_1 \times I_2 \times \ldots \times I_n$, a positive integer **J**, the goal is to factorize **Y** into **N** nonnegative matrices $A^{(n)}$, where is the nth component matrix. These component matrices have dimensions $I_n \times J$and represent the common (or loading) factors.

The tensor **Y** is approximated as a sum of rank-1 tensors formed by the outer product of the vectors from each component matrix $A^{(n)}$. This is mathematically denoted as:

$$\overline{Y} + E = \sum_{j=1}^{J} a_j^{(1)} \circ a_j^{(2)} \circ \ldots \circ a_j^{(n)}$$

where ∘ represents the outer product $\overline{Y}$, is the approximation of **Y**, and **E** is the tensor of residuals or approximation error. The vectors $a_j^{(n)}$ from the component, matrices are constrained to have a unit norm (specifically, the Euclidean norm or L2 norm). This is done to avoid trivial solutions and to ensure uniqueness in the decomposition. The constraint is represented as:

$$\| a_j^{(n)} \|_2 = 1$$

This is the method used to represent the original data by a lower-dimensional approximation obtained via matrix or tensor (multiway array) factorizations or decompositions. Factorization approximates the original tensor. (Fan et al., 2014)

Thera are two methods of updating this algorithm, Alternating Least Squares (ALS), and Hierarchical Alternating Least Squares (HALS). The objective of ALS is to minimize the reconstruction error by iteratively updating each factor matrix while keeping the others fixed. And HALS algorithm is an extension of ALS designed to achieve faster convergence. Unlike ALS, which updates entire factor matrices one at a time, HALS simultaneously updates subsets of the factor matrices. (Huy Phan & Cichocki, 2011)

2.1.2. The non-negative Tucker model

The goal of Non-negative Tucker Decomposition is to find the core tensor core and factor matrices that minimize the difference between the original tensor and the reconstructed tensor obtained by multiplying the core tensor with the outer product of the transposed factor matrices. (Jung et al., 2021)

If we take same original tensor **X** of Nth order with dimensions $I_1 \times I_2 \times \ldots \times I_n$. The Tucker model would decompose this tensor as a core tensor multiplied by factor matrices alone for each dimension. It can be represented as:

$$X \approx G \times x_1 A^{(1)} \times x_2 A^{(2)} \ldots \times x_n A^{(N)}$$

G would represent the core tensor of dimensions. $J_1 \times J_2 \times \ldots \times J_n$. $A^{(n)}$are the factor matrixes of dimension and x_n is the mode **n** product of a tensor with a matrix.

The non-negative constraint is represented by $A^{(n)} \geq 0$ for all **n.**

The core tensor **G** captures the interactions between the different components from each mode. Unlike the PARAFAC model, where the interactions are simplified and directly interpreted, the Tucker model allows more flexibility and can capture more complex relationships. (Asif Malik & Becker, n.d.)

NTF's ability to extract hidden patterns from high-dimensional data makes it a powerful tool in various fields. From biomedical informatics and signal processing to computer vision and data mining, NTF has proven to be versatile and effective in pattern recognition tasks.

Welling and Weber (2001) showed that NTF can identify patterns in muscle activation signals, providing insights into muscle functions and aiding in diagnosing neuromuscular disorders. (Welling & Weber, n.d.) Similarly, Mørup et al. (2008) applied sparse NTF to brain imaging data, revealing significant neural activity patterns and enhancing the understanding of brain functions. (Mørup et al., n.d.)

It is also a proven method for signal processing, particularly in source separation tasks. Yoshii et al. (2013) utilized NTF to separate mixed audio signals, effectively isolating individual sound sources from a composite signal. (Liutkus et al., 2013)

NTF has numerous applications across various fields, but this paper focuses on extracting mobility patterns. The following section delves into research on travel diaries, demonstrating how analyzing them through tensor factorization can benefit public transportation network planning.

2.2. TRAVEL DIARIES

Travel diaries are used as one of the proxies for getting insights into the travel behavior of individuals and groups. These diaries serve as comprehensive records documenting an individual's or a group's travel patterns over a certain period. It provides valuable data for various analyses in urban and transport planning (Prelipcean et al., 2018)

An in-depth activity-travel log provides crucial insights into the daily fluctuations of an individual's patterns and requirements for engaging in activities and travel. This is pivotal for analyses in urban and transport planning. Such a log chronicles details of a person's trips on any given day, encompassing travel duration, choice of transportation modes, preferred

routes to destinations, destinations themselves, and the reasons for making the trips. By gathering this data, researchers can delve into the motivations and inner workings of how individuals make choices about their activities and travel, and how these choices shift across different temporal and spatial settings.

This deepened comprehension can then be synthesized to forecast the impacts of new transit policies or alterations to transport infrastructure, or simply to grasp the ebbs and flows of transport dynamics within specific areas of study. (Prelipcean et al., 2015)

Recent studies have shown a preference for gathering travel diaries through GPS systems, mobile phones, or smart card technology, making such diaries easily available for analysis. (Prelipcean et al., 2017) In contrast, this approach involved data collection through conventional surveys, which enhanced the complexity of the research by incorporating extra details such as the mode of transportation, the purpose of the trip, and demographic information. This additional knowledge can give more insights into individual movement and a better picture of the usage of public transportation networks.

In many studies, the method of using travel diaries to extract movement patterns with Non-negative Tensor Factorization (NTF) has been proven successful.

For instance, Halyal et al. (2022) applied non-negative tensor decomposition to analyze public transit data, uncovering hidden patterns in passenger travel behavior and forecasting transit demand. This study demonstrated the utility of NTF in understanding and forecasting public transit demands based on passenger data from automatic passenger counters (APC) and AVL data. (Angadi et al., 2023).

Another notable study by Tang et al. (2020) employed NCP decomposition on smart card data from Shenzhen's metro system to identify stable spatio-temporal travel patterns. This study highlighted how NTF can reveal intricate travel behaviors and support urban planning and the operation of metro networks by understanding departure and arrival patterns linked to specific urban functions (Tang et al., 2020). This would be very beneficial information for urban network planners. This would be very.

We also found studies that focus on public bike usage. Chen et al. (2019) studied bike-sharing systems using tensor factorization. This research focused on the large-scale dynamics of bike-sharing data, revealing significant patterns in user behavior and station usage across different times and locations. The study applied NTF to collaboratively infer temporal activity preferences collaboratively, helping in understanding how bike-sharing systems are utilized in urban environments and optimizing bike distribution and station placement. Optimizing bike stations and lanes is also one of the ways of lowering car usage in big cities. (Wang et al., 2022), (García-Palomares et al., 2012)

Focusing on bus rides is easier in the cities that have implemented smart card systems. We found that both inner-city and interconnecting bus rides exploration brings benefits to urban planners. Studies note that NTF uncovers hidden patterns and provides insight into effective and efficient public transportation systems. It allows exploration of different region connections and how behaviors change over time. (Z. Wang et al., 2023), (Haghighat et al., 2020)

Understanding the impact of behavioral changes on travel patterns is crucial for effective urban planning. Fan et al. (2014) highlighted how significant events, such as the Great East Japan Earthquake, can drastically alter mobility patterns. Their study showed that post-disaster, people's commuting patterns changed significantly, with many opting to stay at home or relocate to safer areas. This behavioral shift was quantitatively analyzed using NTF, revealing the resilience and adaptability of urban populations in response to disasters (Fan et al., 2014). The analysis of post-disaster mobility patterns helps in planning for emergency responses and long-term urban recovery strategies.

Another beneficial approach to this exploration would be integration of travel diary data with other data sources, such as demographic information and land use data, which can further enhance the understanding of travel behaviors. This holistic approach allows for a more comprehensive analysis of how different factors influence travel patterns and how changes in one aspect (e.g., a new transportation policy or infrastructure development) can impact overall mobility. For example, integrating travel diary data with demographic and land use data has shown that land use characteristics and socio-demographic factors significantly influence travel behavior and activity spaces (Sharmeen 2021). Such integration helps transport planners and policymakers to design more effective and equitable transportation systems by considering the diverse needs and behaviors of different population groups (van Wee, 2018).

By leveraging modern data collection technologies and integrating multiple data sources, researchers can perform more detailed and sophisticated analyses, leading to better-informed decisions in urban and transport planning. The use of smart card data in metro systems, for example, provides continuous and precise records of passenger movements, which can be analyzed to identify peak travel times, popular routes, and potential bottlenecks in the system (Mützel & Scheiner, 2022). Additionally, integrating land use data helps in understanding how different areas of a city are connected and how changes in land use can affect travel patterns (Sharmeen, 2021).

Travel diaries have been shown as an effective source of information. They help in understanding travel behaviors and informing urban and transport planning. The application of NTF and other advanced analytical methods to travel diary data allows for the extraction of meaningful patterns and insights that can improve the efficiency of transportation systems. By leveraging modern data collection technologies and integrating multiple data sources,

researchers can provide robust analyses that support sustainable and informed urban development.

In this paper, we want to follow the analysis of leveraging NTF methos for extraction of movement patterns. The aim is to compare two algorithms we found most used in studies, the non-negative PARAFAC and non-negative Tucker algorithm. The goal is to see which of the algorithms will give better quantitative but also qualitative results.

3. METHODOLOGY

In this research, we aim to uncover hidden patterns of urban mobility using Non-negative Tensor Factorization (NTF) applied to travel diary data. This section outlines the dataset, preprocessing steps, tensor construction, and the algorithms used to extract movement patterns.

The framework for this study involves several stages: data understanding, data preparation, tensor construction, algorithm application, and pattern extraction. Figure 1 depicts a conceptual model illustrating the flow from raw data to extracted patterns.

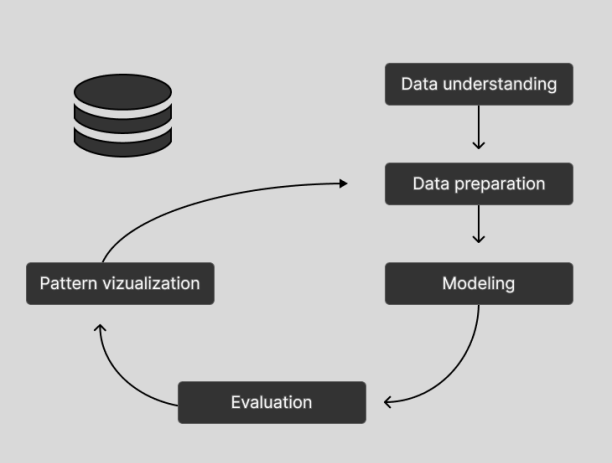

Figure 1 - Flow of experiment

The steps will be more closely described in the following text, but overall, they include data understanding, data preparation, modeling, evaluation of the models, and pattern visualization.

3.1. DATA UNDERSTANDING

The dataset used in this research is collected by The New York City Department of Transportation. They conduct a Citywide Mobility Survey once per year, the aim is to evaluate travel behavior, preferences, and attitude of New York City residents. In 2019. CMS got 3346 citizens to participate and collected all trips made by them, from May 2019 through June 2019. The survey can be answered via smartphone, online, or through the call center. The same questionnaire was used for all participants, regardless of the method of participation.

The CMS has been collecting two primary types of data: demographic information and travel diary information. In total, there were 85459 trips recorded in this study, and each trip is detailed through 79 descriptive points. (*Open_Data_Dictionary_CMS_Trip_Survey_2019v2*, n.d.).

As part of the data understanding process, Exploratory data analysis was also conducted, and the key insights of that analysis will be shown in the next chapter.

3.2. DATA PREPARATION

From the original dataset, we take 3 dimensions of data: day of the week, purpose of the trip, and mode of transportation. Originally the research intended to explore a higher number of dimensions, but because of the way the data was collected some of the dimensions were not providing valuable information for this experiment.

After analysis and cleaning of the original dataset, we conducted the approach for our experiment of pattern extraction, using only three of the original columns as three dimensions of the tensor that we will be building. Those dimensions are the day of the week, mode of transportation, and purpose of the trip.

The original experiment was planned with more than three dimensions, but after testing several other columns, we decarded some for several reasons. For example, the dataset includes origin and destination location, and they were supposed to be part of the results, but they are not included for two main reasons:

1. The values are recorded as neighborhoods, a specific part of a neighborhood, or one of New York's airports (JFK, and LGA). There is no record of specific stations on which the trip has started/ended.

2. In The vast majority of recorded trips (> 80%), the origin point, and destination point have been the same.

Also, the original dataset includes more modes of transportation than we included in our experiment. The experiment explored trips by Subway, Bus, Bike, Ferry, Commute rail, and other categories. The original survey also included the Vehicle category and walk category. Because of unbalanced records for those two categories, they were not included in the final experiment. As is shown in Figure 2. the usage of Vehicles and Walking is double the usage of all public transportation modes combined. These led to patterns that highly favored these two modes, even after algorithms were parametrized for unbalanced data. Also, they are not part of the public transportation network, which is the main topic of this exploratory experiment.

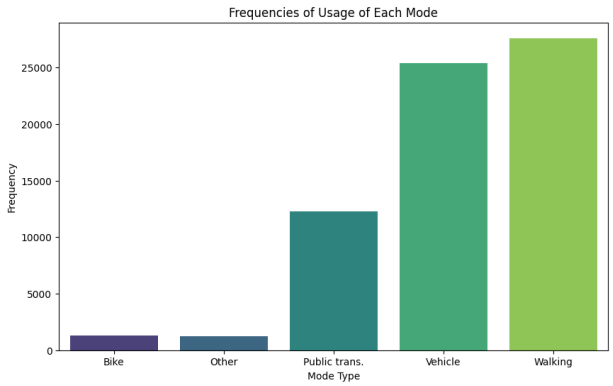

Figure 2 - Frequencies of Usage of Each Mode

To conduct three dimensions, we are going to explore are

- **Day of the Week:** This dimension helps in understanding temporal patterns.
- **Mode of Transportation:** Essential for identifying the use of different modes of transport.
- **Purpose of the Trip:** Provides context for the trips, differentiating between work commutes, shopping, and leisure activities.

The dataset needs to be transformed into a multidimensional array, a crucial step for it to be suitable for tensor decomposition. Tensor decomposition techniques require data to be structured in a multiway array format, commonly known as a tensor. In this context, a tensor is a generalization of a matrix to higher dimensions, allowing for the analysis of data that varies across multiple axes or modes. (Shashua & Hazan, n.d.)

For our analysis, we specifically focus on three dimensions derived from the original dataset: the day of the week, the mode of transportation, and the purpose of the trip. Each of these dimensions represents a column in the dataset, and the unique values within each column determine the size of the tensor along that dimension. The day of the week has 7 unique values (Monday to Sunday), the mode of transportation has 6 unique values (e.g., subway, bus, bike, etc.), and the purpose of the trip has 10 unique values (e.g., work, home, shop, etc.). The resulting tensor is a three-dimensional array of 7 x 6 x 10.

3.3. ALGORITHM

Choosing the right algorithm for this experiment focused on two main objectives: extracting non-negative patterns and ensuring clear interpretability of patterns. The literature mentions two main algorithms that satisfy both requirements.

Non-negative PARAFAC algorithm decomposes multiple tensors into a sum of rank one tensors, with the constraint of non-negativity. This approach was selected for its ability to reveal underlying patterns while maintaining data interpretability. Two versions of the PARAFAC algorithm were considered for this study: ALS (Alternating Least Squares) and HALS (Hierarchical Alternating Least Squares). Both approaches aim to decompose a tensor into a sum of rank-one tensors, but they differ significantly in their methods for updating component matrices during the factorization process.

The ALS method is a straightforward and widely used approach for tensor decomposition. In this method, one factor matrix is updated at a time while keeping all other factor matrices fixed until convergence is achieved. Specifically, the ALS algorithm cyclically updates each matrix by solving the least squares problem. This process iteratively minimizes the reconstruction error of the tensor. ALS is relatively easy to implement and understand, which contributes to its popularity in various applications of tensor decomposition. A key limitation of ALS is its potentially slow convergence, particularly when dealing with large-scale tensors

or when the initial values are far from the optimal solution. This can result in longer computation times and increased resource consumption. (Huy Phan & Cichocki, 2011)

The HALS algorithm, also known as Hierarchical Alternating Least Squares, offers an enhancement over the traditional ALS method. HALS updates a subset of factor matrices simultaneously, using a hierarchical approach. This means that instead of updating one-factor matrix at a time, HALS updates groups of elements within the factor matrices, which can lead to faster convergence. This model typically achieves convergence more rapidly than ALS due to its ability to handle the optimization problem more efficiently. By updating multiple components concurrently, HALS reduces the number of iterations required to reach an optimal solution, making it particularly suitable for large-scale problems. (Huy Phan & Cichocki, 2011)

Figure 3 illustrates the PARAFAC decomposition process, showing the full mobility dataset decomposed into smaller components corresponding to the three dimensions.

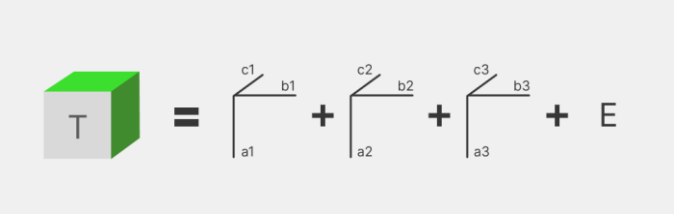

Figure 3 - Non - negative PARAFAC decomposition

The Non-negative Tucker Decomposition (NTD) algorithm was also explored. NTD decomposes a multi-dimensional tensor into a core tensor and factor matrices, capturing interactions between modes in a reduced-dimensional space (Jung et al., 2021). This method compresses the original data while retaining significant patterns.

Figure 4 shows the Tucker decomposition, where the original dataset is reconstructed into a core tensor and associated factor matrices (A, B, C).

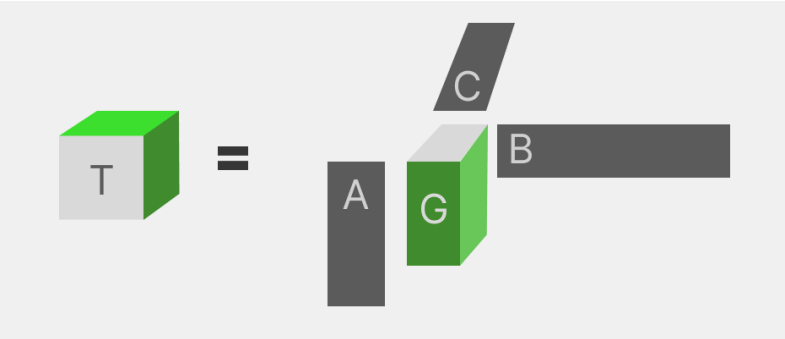

Figure 4 - Non - negative Tucker decomposition

The algorithms were implemented using the Python Tensorly library, which provides robust tools for tensor decomposition. (Shashua & Hazan, n.d.) Three approaches were tested:

- Non-negative PARAFAC (ALS)
- Non-negative PARAFAC (HALS)
- Non-negative Tucker

A critical parameter in these algorithms is the *rank*, which determines the dimensionality of the extracted patterns. To identify the optimal *rank*, we evaluated performance across different ranks, balancing reconstruction error and explained variance. The chosen rank aimed to capture the most significant patterns while minimizing noise. Other than rank, other parameters have been set following library suggestions for optimization.

n_itter_max, used in all algorithms, was set to 100. This parameter specifies the maximum number of iterations for NTF algorithm.

init parameter was set to 'svd', for all the algorithms. Init determines initialization method. Specifically, SVD means that algorithm will initialize the factor matrices using Singular Value Decomposition.

normalize_factors parameter was set to True. This parameter is only available when using PARAFAC algorithms. It controls the scale of each factor which leads to more interpretable results. Important to set in case of unbalanced datasets.

In the next section, we will discuss the results of testing each model on constructed tensors and final patterns extracted from the dataset.

4. RESULTS AND DISCUSSION

The results section presents the outcomes of our comprehensive analysis, focusing on the patterns and insights extracted from the travel diary data. This section is divided into several parts, each detailing the findings from different analytical approaches, including Exploratory Data Analysis (EDA), the application of Non-negative PARAFAC models, and the Non-negative Tucker Decomposition.

4.1. EXPLORATORY DATA ANALYSIS

The exploratory data analysis (EDA) phase was essential for understanding the dataset's general characteristics and laying the groundwork for subsequent pattern extraction. The analysis focused on trip frequencies based on three main categories: the day of the week, mode of transportation, and trip purpose. Each of these categories provided valuable insights into the overall travel behaviors in New York City.

The analysis revealed that most trips occurred on Mondays, and trips were more frequent throughout the weekdays than on weekends. Figure 5 illustrates the distribution of trips across the days of the week. Monday emerged as the most traveled day, likely due to the workweek's start, while the other weekdays also showed substantial travel activity. This trend suggests a consistent pattern of weekday commuting, with weekends exhibiting slightly lower travel frequencies.

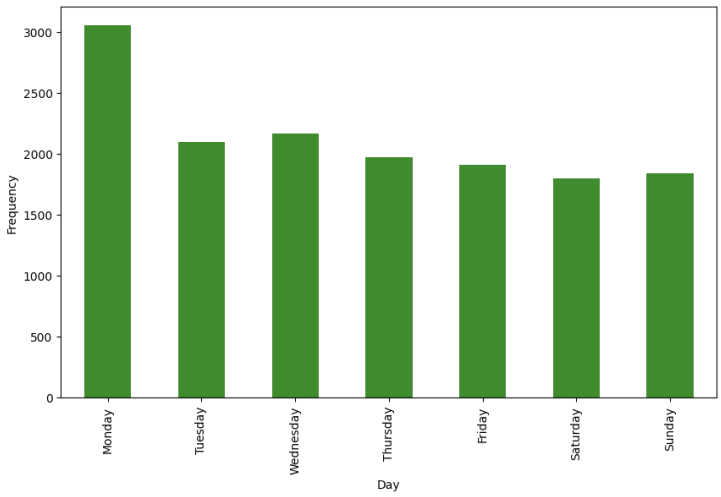

Figure 5 - Frequency of trips per day of the week

Figure 6 displays the frequency of trips by different modes of transportation. The subway system dominated as the primary mode of transport, with more than double the number of trips compared to the second most used mode, which was the bus. This dominance of subway usage underscores its critical role in the city's daily transportation network. The usage of other modes, such as buses, bikes, and ferries, is less frequent, but it still can indicate a diverse and multimodal transportation system.

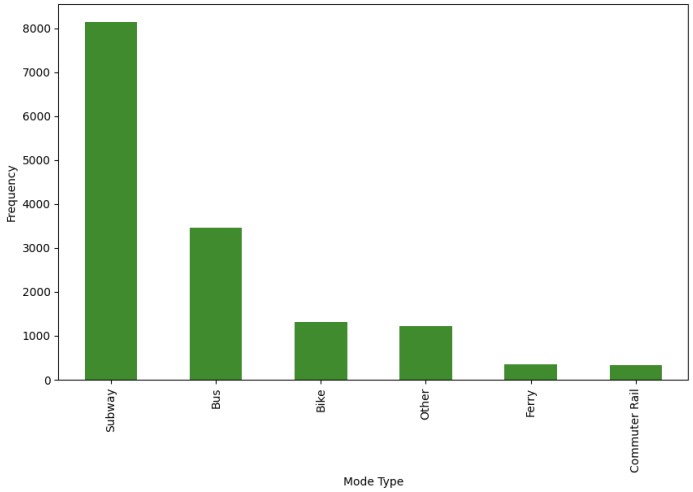

Figure 6 - Frequency of trips per mode of transportation

Figure 7 illustrates the frequency of trips based on their purpose. Commuting to work and home were the predominant purposes, reflecting the routine nature of daily travel. Other significant categories included shopping, social, and recreational trips. The distribution of trip purposes highlighted the varied nature of travel needs within the city, ranging from essential commutes to discretionary activities.

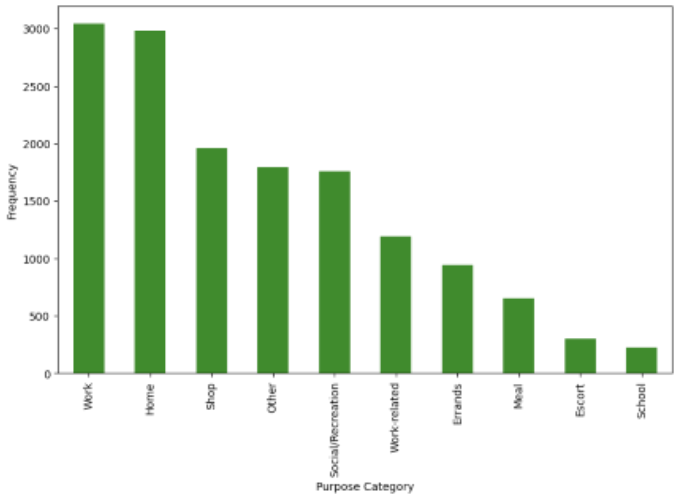

Figure 7 - Frequency of trips per trip purpose

Given the subway's dominance, a more detailed exploration of subway trips was conducted. The distribution of subway trips by the day of the week mirrored the overall trip distribution, with Monday being the most traveled day and similar frequencies observed across other weekdays. This further emphasized the subway's role in daily commuting patterns. Additionally, the distribution of trip purposes for subway trips showed a notable drop in home commutes and a higher frequency of undefined trips compared to the overall distribution. This suggests that subway trips might include a broader range of activities beyond commuting.

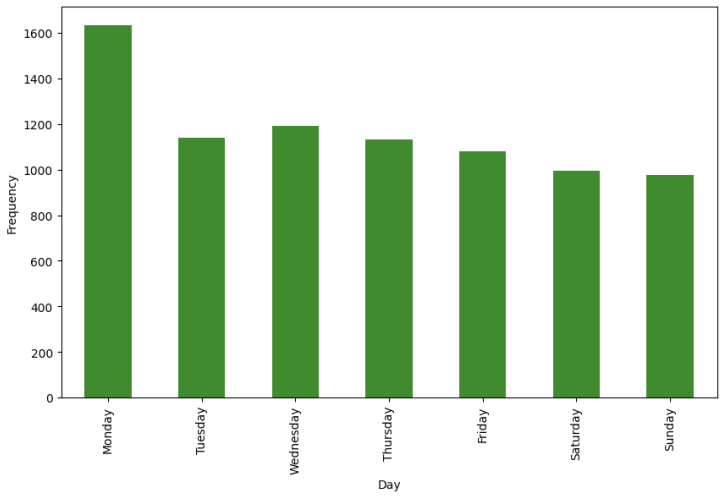

Figure 8 - Day distribution of Subway trips

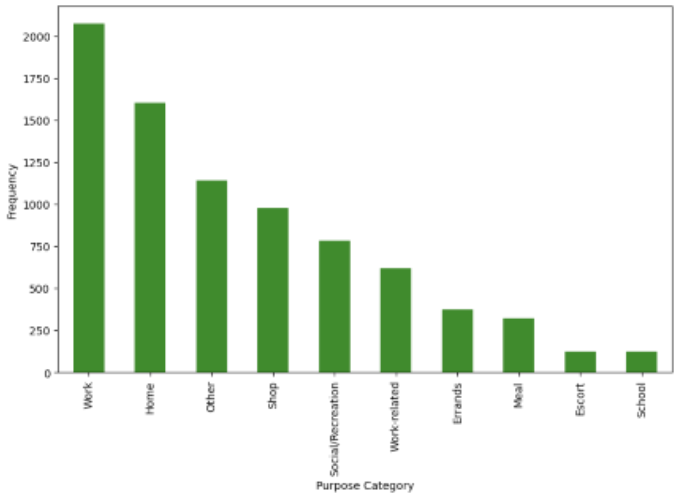

Figure 9 - Purpose distribution of Subway trips

Based on the EDA findings, several expectations were set for the extracted patterns. Given its dominance, it was anticipated that the patterns would emphasize trips taken by the subway. The bus was also expected to feature prominently due to its significant usage. Monday was predicted to be highlighted as a highly traveled day, especially for subway trips, reflecting its overall trip frequency. Additionally, the results were expected to show patterns for commutes to work and home, with potential differentiation in the mode of transport for home commutes. The patterns were also anticipated to capture shopping and social/recreational trips, reflecting the diversity of trip purposes observed in the EDA.

The insights gained from this phase played a key role in shaping the next steps of our analysis. They helped us focus our efforts on the most important aspects and made sure that the patterns we found were clear and made sense.

4.2. ALGORITHM OPTIMIZATION

After initial exploration and preparation of data for usage with tensor factorization, we started optimizing models. Optimizing the algorithms was a critical step in our research to ensure that the extracted patterns were meaningful and interpretable. The default parameters used are explained in the Methodology section, but we will continue optimizing the number of ranks for each model.

We employed the elbow method to determine each algorithm's optimal rank. This involved balancing reconstruction errors and explaining variance. Figures 10, 11, and 12 illustrate the performance of the three algorithms (Non-negative PARAFAC ALS, Non-negative PARAFAC HALS, and Non-negative Tucker) across different ranks.

The reconstruction error is calculating the Frobenius norm of the difference between the original tensor and the reconstructed tensor. This norm measures how much the reconstructed tensor deviates from the original tensor, indicating the accuracy of the decomposition. The explained variance number measures of how much of the original tensor's variance is captured by the decomposition algorithm.

Figure 2 shows a graph for Non-negative PARAFAC ALS. It displays the reconstruction error and explains variance across various ranks. The elbow point, where the rate of decrease in error slows down, suggests that a rank of 5 provides an optimal balance between error and interpretability.

Figure 3. shows graph for Non-negative PARAFAC HALS, the analysis indicates that a rank of 5 is also optimal. This version of the algorithm shows slightly better performance in terms of reconstruction error compared to ALS that suggests that HALS method, in this case, can capture the data structure more efficiently.

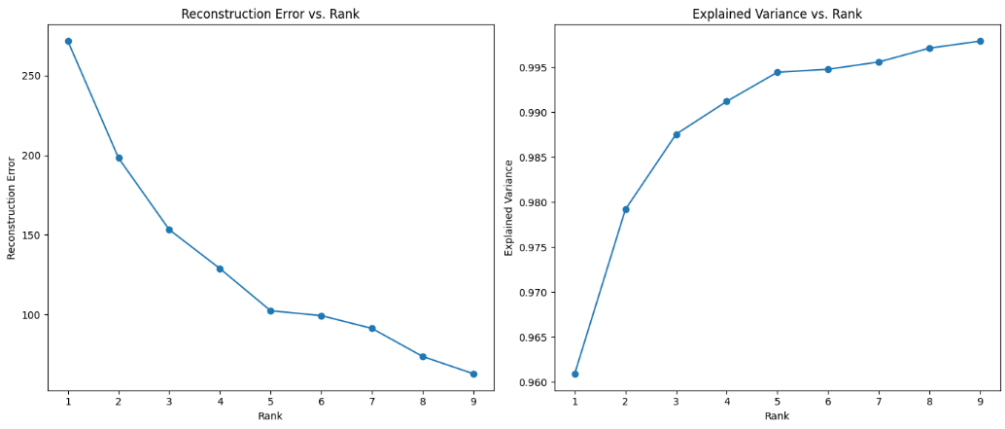

Figure 10 - Algorithm 1

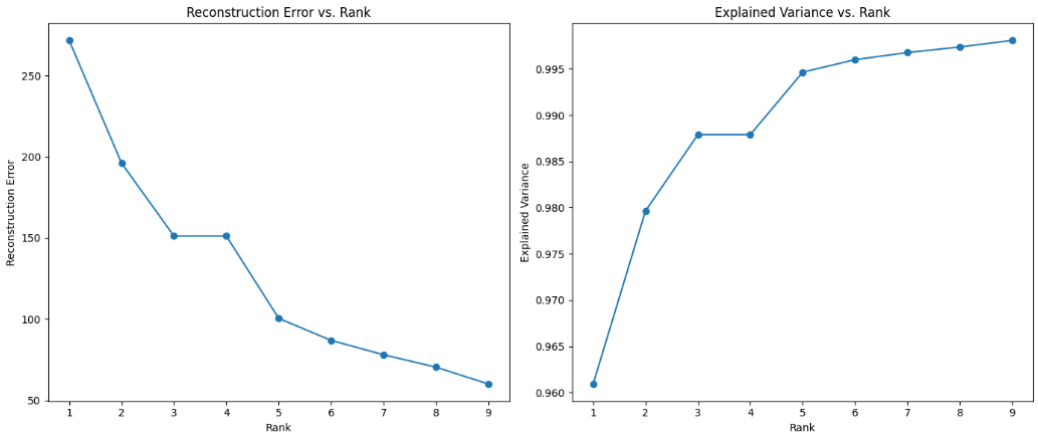

Figure 11 - Algorithm 2

Finding the optimal rank for Tucker algorithm was more complex because Tucker's composition can operate with different ranks for each dimension. In Figure 7 you can see the balance between error and variance explained if we restrict rank to be the same value. Elbov method, in this case, suggests that rank 6 would be optimal.

To use the algorithm's full potential, we alternate between unique rank values for every tensor dimension. The maximal rank value is set according to the tensor shape. For our tensor in the shape of 7 x 6 x 10, we set the maximal rank value of 7 for the first dimension and 6 for the second. And 10 for the third.

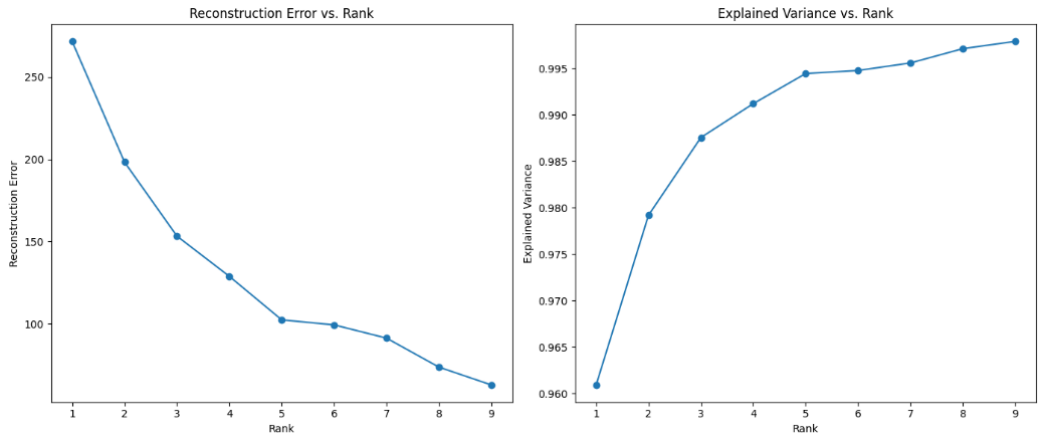

Figure 12 - Algorithm 3

After finding the ideal rank for each of the algorithms, we compared the results of all 4 models of them based on reconstruction error and Explained Variance. The differences are shown in Table 1. After inspecting the results, we can see that the variance explained is high in all four models, going from 99.5% to 99.9%. Those differences are insignificant. Reconstruction error varies more between models: 79.02 for the Tucker model (rank – 7, 6, 8) to 102.37 for the ALS PARFAC model. As differences between variances explained are insignificant, we well take that

model with smaller reconstruction error has better quantitative results, so the Tucker model (rank – 7, 6, 8) would be the best, followed by second Tucker model, than PARFAC Hals, and at the end PARFAC Als.

Table 1 - Comparison of Algorithm Performance Metrics

Algorithm	Rank	Reconstruction Error	Variance explained
PARFAC	5	102.37	0.999
PARFAC Hals	5	100.51	0.994
Tucker	[6, 6, 6]	94.31	0.995
Tucker	[7, 6, 8]	79.02	0.996

To extract interpretable patterns effectively, finding an optimal balance between descriptive and quantifiable measures is crucial. Although the Tucker algorithm gives better mathematical results with varying ranks. The inconsistency between the ranks leads to harder pattern extraction. To choose an algorithm that will best represent this data, we must return to the experiment's main goals. All the approaches follow nonnegativity, but the second requirement is interpretability. The pattern extraction interoperability highly depends on the rank number, and it is most coherent when that number is the same for all the dimensions.

For that reason, we will not look at the results of the Tucker model with inconsistent ranks, as the patterns will not be clear. In the next section, we will visualize and compare extracted patterns for two next best models from Table 1.

4.3. EXTRACTED PATTERNS

Firstly, we will explore the results of the Tucker model with rank 6 for each dimension. We applied Non-negative Tucker Decomposition to the travel diary data using the optimal parameters identified. The model had a rank value set to 6 for all three dimensions. This process resulted in six distinct movement patterns, each providing insights into the relationships between the day of the week, the mode of transportation, and the trip's purpose.

To better understand the extracted patterns, we created heatmaps for each dimension (day of the week, mode of transportation, and trip purpose) across the six components of the decomposed tensor. These heatmaps illustrate the significance of each dimension within the identified patterns, providing a visual representation of the complex interactions between different travel behaviors. Heath maps are shown in Figure 13.

From the heatmaps, we extracted patterns of movement. Table 2 shows the most significant components of each of the six patterns. These are not the only components but the main drivers of movement for that pattern.

From the table, we can see that Monday is predominant in half of the patterns, which corresponds with exploratory data analysis showing a higher Frequency of trips made on Monday. Half of the patterns are also connected to trips done by Subway; again, that was expected from data analysis. For the purpose dimension, Work has the most significance.

Table 2 - Extracted travel patterns for Tucker model

Pattern	Day category	Mode category	Purpose category
1	Saturday	Subway	Work
2	Wednesday, Thursday	Subway, Other	Work, Other
3	Monday, Friday, Saturday	Bus, Bike	Shopping, Home
4	Sunday, Tuesday	Other	Work, Errands
5	Monday, Wednesday	Subway, Ferry	Work, Home
6	Monday, Tuesday	Bus	Social/Recreational

Pattern 1

This pattern suggests significant use of the subway for work commutes, particularly at the start of the week and on Saturdays. This indicates possible shifts in work schedules or recreational activities that require subway travel on Saturdays.

Pattern 2

This pattern highlights mid-week travel behavior, mostly done by Subway, with a mix of Work-related and Other unspecified activities. This may indicate mid-week meetings, flexible work arrangements, or a variety of mid-week errands that require travel. The dual nature of the trips (work and other purposes) suggests a diverse use of the subway system.

Pattern 3

This pattern reveals a higher usage of Buses and Bikes at the start and end of the workweek and on Saturdays. These trips are primarily for shopping and returning home, suggesting these

days are popular for running errands and engaging in leisure activities. The use of buses and bikes indicates a preference for these modes for shorter, more localized trips.

Pattern 4

This pattern indicates a varied transportation mode usage on Sundays and Tuesdays, with a mix of work-related and other errands. The diversity in transportation modes suggests that these days involve a range of activities and trip purposes, reflecting the flexibility in how people travel these days. The presence of 'Other' modes indicates less common or multiple forms of transportation being utilized.

Pattern 5

This pattern shows a high frequency of subway use on Mondays and Wednesdays, along with notable ferry usage. This highlights these days as significant for commuting, possibly linked to weekly ferry schedules or specific work locations near ferry routes. The combined use of subways and ferries underscores the importance of multimodal transportation networks in urban commutes.

Pattern 6

This pattern points to high bus usage on Mondays and Tuesdays for social and recreational activities, indicating these days are popular for non-work-related travel. The use of buses for social outings and recreational trips suggests a preference for this mode for less urgent, flexible travel.

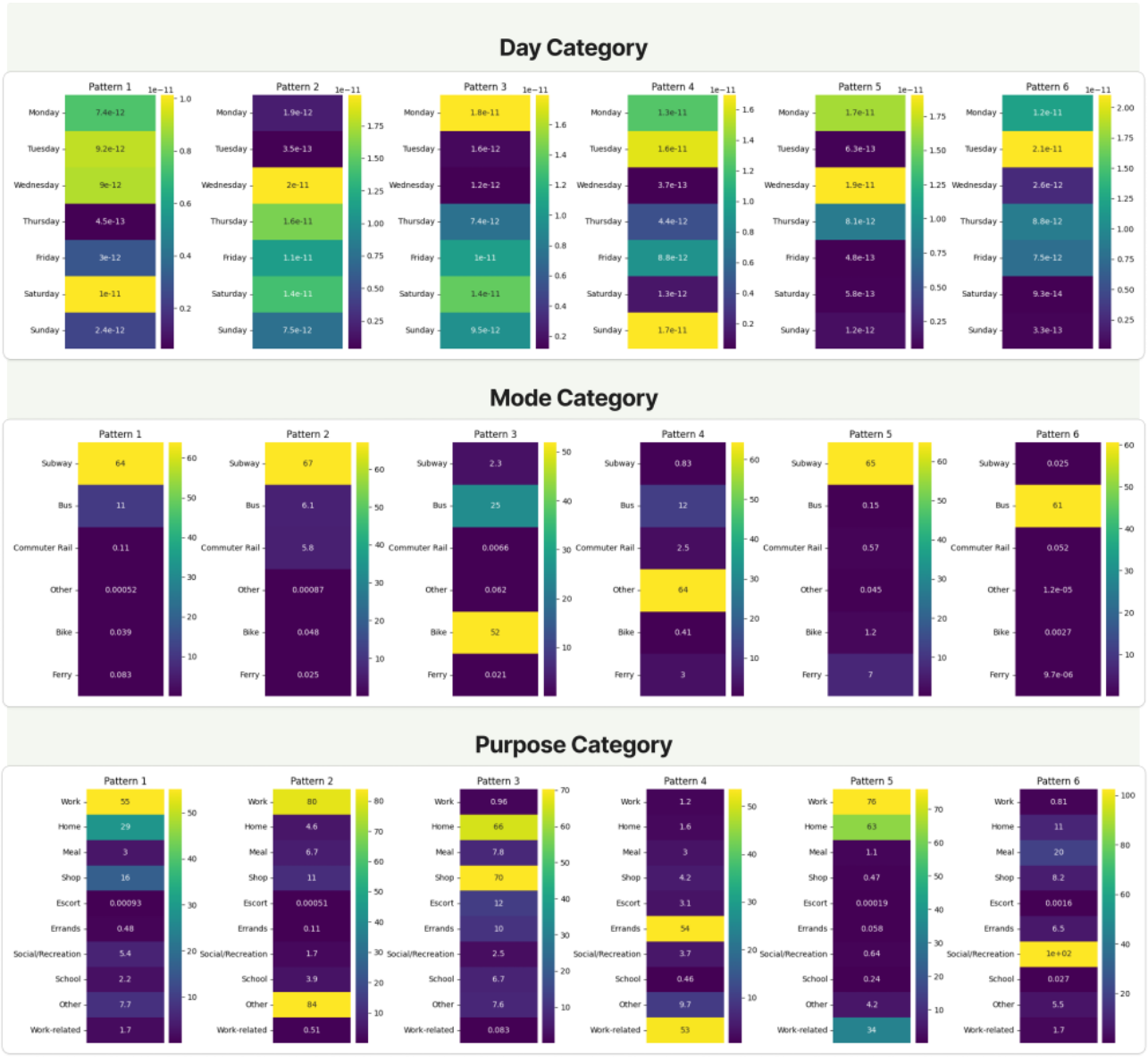

Figure 13 - Extracted travel patterns for Tucker model

To better understand patterns and overall movements in the dataset, the previously plotted results will be compared with the results of the second model. That will be PARAFAC model HALS, with rank set to 5. The visualization shows similar patterns to those we described previously. The heatmaps in Figure 14 show all the patterns for each dimension.

First, we will describe patterns extracted with this model and then compare them with previous models' findings.

Again, Table 3. It shows the most significant components of extracted patterns, similar to the results of the first model. Monday is the most featured day of the week, and the subway is the most used vehicle. The purpose dimension shows less significance in the Work category.

Table 3 - Extracted travel patterns for PARAFAC model

Pattern	Day category	Mode category	Purpose category
1	Monday	Subway	Work, Home
2	Tuesday	Subway	Other
3	Monday	Bus	Home, Shop
4	Monday, Thursday	Subway	Social/Recreational
5	Sunday, Friday	Other	Work-related, Errands

Pattern 1

This pattern indicates a high transportation usage on weekdays, particularly peaking on Monday and Thursday, and the usage significantly drops on weekends, with Saturday being the lowest. The dominant mode of transport is the Subway, followed by Buses. The primary purpose for travel is Work, with Home being the second highest. This pattern suggests a routine workweek commute with predominant reliance on subways and a clear focus on work-related travel.

Pattern 2

This pattern shows peak transportation usage on Tuesday and Thursday, and the subway is the dominant mode of transport. The Other, undefined purpose is shown as the highest in this pattern, but there are also many trips to Work.

Pattern 3

This pattern highlights higher transportation usage on Monday and Tuesday. Buses are used more than other modes of transportation and shopping is the primary purpose of travel, followed by trips Home. This pattern shows trips done by Bus. Busses are mostly used at the beginning of the week for non-work-related trips.

Pattern 4

This pattern indicates consistent and relatively high transportation usage from Monday to Thursday, peaking on Monday. The Subway remains the primary mode of transport, but there is noticeable usage of Other mode, and also of Bikes. The primary purpose of travel is Social/Recreational, but there is also a significant number of trips back home.

Pattern 5

This pattern shows peak transportation usage on Sunday, with higher number also on Friday, and Wednesday. The dominant mode of transport is classified as 'Other,' indicating we don't have information about the exact mode of transport for these trips. The primary purpose of travel is work-related, followed by Errands trips. This pattern suggests diverse transportation modes with a mix of work-related and other errands, reflecting flexibility in travel activities and purposes.

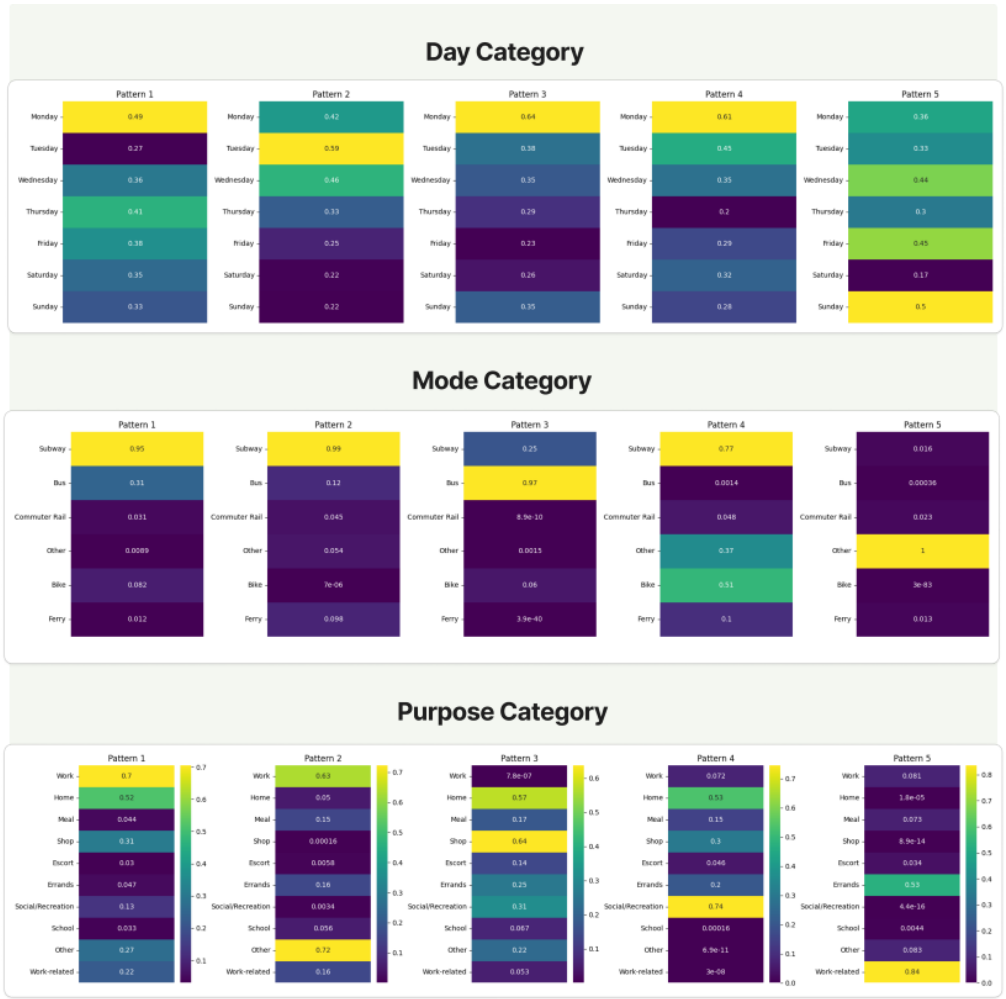

Figure 14 - Extracted travel patterns for PARAFAC model

4.4. SUMMARY

The exploratory data analysis (EDA) aimed to understand the dataset's general characteristics, focusing on trip frequencies by the day of the week, mode of transportation, and trip purpose. Key findings included that the highest number of trips occurred on Mondays, with generally high frequencies throughout the weekdays. Subway trips were the most frequent, significantly outnumbering other modes such as buses, and the primary purposes of trips were commuting to work and home. The extracted patterns from the tensor decomposition methods aligned closely with these EDA insights but provided more granular details.

The patterns provided a deeper understanding of travel behaviors, highlighting specific days and modes of transport that were not as apparent in the initial EDA.

Two different tensor decomposition models were used, the Non-negative Tucker Decomposition and the Non-negative PARAFAC model with HALS, to extract patterns from the travel data and both models had their strengths but revealed similar but distinct insights. The Non-negative Tucker Decomposition, with an optimal rank set to 6 for all dimensions, provided detailed patterns with high diversity in daily travel behaviors. It identified six distinct patterns, each combining different days, modes of transportation, and trip purposes, highlighting more complex interactions and being more effective in showing a wide range of travel activities across different days.

On the other hand, the PARAFAC model, with an optimal rank set to 5, focused more on weekday commuting patterns, showing heavy subway usage on Mondays and Thursdays. The patterns extracted by this model were simpler and more straightforward, with a clear emphasis on work commutes and primary transportation modes. This model highlighted consistent and regular travel behaviors. That made the results easier to interpret but less detailed in terms of daily variations.

The main difference between the two models lies in the distribution of transportation modes and trip purposes: the Tucker model provided more diverse daily travel insights, whereas the PARAFAC model emphasized weekday commuting patterns. We can see examples in comparisons of first and third patterns of both models.

Tucker's Pattern 1 indicated significant subway usage on Saturdays for work commutes, suggesting shifts in work schedules or recreational activities on weekends. In contrast, PARAFAC's Pattern 1 showed high subway usage on Mondays for work and home commutes, highlighting a routine weekday commuting pattern with a clear focus on the start of the workweek.

Tucker's Pattern 3 revealed higher usage of buses and bikes on Mondays, Fridays, and Saturdays for shopping and home trips, indicating these days are popular for running errands and engaging in leisure activities. On the other hand, PARAFAC's Pattern 3 focused on higher transportation usage by bus on Mondays and Tuesdays for shopping and home trips, suggesting the beginning of the week is primarily for non-work-related activities using buses.

Overall, the comprehensive analysis demonstrated that subways were the primary mode of transportation, especially for work, and work-related commutes, aligning with EDA findings. The Tucker model uncovered complex travel behaviors, showing significant variations across different days and purposes, while the PARAFAC model, though less detailed, provided clear and coherent patterns of weekday commuting behaviors, emphasizing the regular use of subways and buses for work and home trips. These insights are crucial for urban transportation planning. They highlight the need for efficient multimodal transport networks and address varying travel demands throughout the week.

Tensor decomposition techniques extracted meaningful travel patterns, providing valuable insights for transportation planning and policymaking. The detailed patterns revealed by the models can help design better transportation services, optimize transit schedules, and improve the overall efficiency of urban mobility systems. These insights are based on the dataset and assume that the data collection truthfully showcases the frequency of each dimension.

The subway system is the most frequently used mode of transportation, particularly for work commutes, underscoring its critical role in the daily transportation network of New York City. Also, patterns show that buses and bikes are commonly used for shopping and recreational activities, especially at the beginning and end of the week, suggesting a preference for these modes for shorter, more localized trips. Then, different days of the week exhibit distinct travel patterns, with specific modes and purposes dominating certain days. This temporal variation indicates the need for flexible transportation planning that accommodates changing travel behaviors. Lastly, including ferries in the commute highlights the importance of water-based transportation for specific travel needs, underscoring the value of multimodal transportation networks in enhancing connectivity and providing efficient travel options.

5. CONCLUSIONS

In this research non-negative tensor factorization was applied to travel diary data, with the goal of discovering hidden patterns and trends in citizens' behavior. The focus of the research was public transportation in New York, by taking advantage of the New York City Department of Transportation's annual citizen mobility survey.

The exploratory data analysis (EDA) established foundational insights into travel frequencies, revealing that weekdays, particularly Mondays, exhibit the highest travel volumes. Subway travel emerged as the predominant mode of transportation, especially for work-related commutes, emphasizing its critical role in the city's daily transportation network. The detailed patterns extracted through NTF extended these findings, offering a granular understanding of travel behaviors across different days and purposes.

Both the Non-Negative Tucker and PARAFAC models were effective in uncovering intricate travel patterns. The Tucker model, with its optimal rank configuration, revealed complex interactions between various dimensions, capturing a wide range of travel activities. It demonstrated significant variations in daily travel behaviors, emphasizing the importance of multimodal transportation networks in addressing diverse travel needs. The PARAFAC model, on the other hand, provided a more straightforward interpretation of weekday commuting patterns, highlighting the consistent use of subways and buses for work and home trips. Both models highlighted the dominance of the subway system for work commutes. Patterns also showed significant use of buses and bikes for shopping and recreational activities, and notable and consistent travel patterns were associated with weekdays. These findings underscore the critical role of public transportation in urban mobility and offer an understandable explanation of how different factors influence travel behaviors.

The results offer practical implications for urban planners and policymakers, supporting data-driven decision-making for sustainable and efficient urban transportation systems. The research sets a foundation for future studies to build upon, particularly with enhanced datasets and extended temporal analyses, to further refine and expand our understanding of urban mobility dynamics.

6. LIMITATIONS & FUTURE WORK

The biggest limitation of this experiment ended up being the initial dataset. The initial survey that collected the data records had different goals from this experiment, so we did not have all the information relevant for finding patterns of movements. This lack of detailed data significantly impacted on our ability to draw comprehensive and accurate conclusions about travel behaviors.

A few suggestions that would help improve the relevance of the results are:

- Obtaining origin and destination data, either as coordinates or station names, to better understand travel routes.

- Recording the timestamps of trip start and end times to analyze temporal patterns more effectively.

- Extending the data collection period. In this case, we only had one month of collected trips, so it would be beneficial to see how movement patterns change across the whole year and identify any seasonal variations.

Another limitation is that there were many initial goals; the original survey was obtaining information about the whole scope of New York citizen's movements. But for a more in-depth analysis, an example of this research would be an analysis of public transportation, the suggestion would be to focus on one aspect of transportation to get more details information for example (time of waiting for transportation, delay time, change mode information) Implementing these suggestions would enhance the quality and depth of results.

BIBLIOGRAPHICAL REFERENCES

Agencia Europea de Medio Ambiente. (n.d.). *Transport and environment report 2022 : digitalisation in the mobility system : challenges and opportunities*. Publications Office of the European Union, 2022.

Angadi, V. S., Halyal, S., & Mulangi, R. H. (2023). Spatiotemporal capacity estimation of bus rapid transit system based on dwell time analysis. *Journal of King Saud University - Engineering Sciences*. https://doi.org/10.1016/j.jksues.2023.10.001

Asif Malik, O., & Becker, S. (n.d.). *Low-Rank Tucker Decomposition of Large Tensors Using TensorSketch*.

Bro, R. (1997). Chemometrics and intelligent laboratory systems Tutorial PARAFAC. Tutorial and applications. In *Chemomemcs and Intelligent Laboratory Systems* (Vol. 38).

Curado, M., Tortosa, L., & Vicent, J. F. (2021). Identifying mobility patterns by means of centrality algorithms in multiplex networks. *Applied Mathematics and Computation*, *406*. https://doi.org/10.1016/j.amc.2021.126269

Engin, Z., van Dijk, J., Lan, T., Longley, P. A., Treleaven, P., Batty, M., & Penn, A. (2020). Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, *9*(2), 140–150. https://doi.org/10.1016/j.jum.2019.12.001

Fabbiani, E., Nesmachnow, S., Toutouh, J., Tchernykh, A., Avetisyan, A., & Radchenko, G. (2018). Analysis of Mobility Patterns for Public Transportation and Bus Stops Relocation. *Programming and Computer Software*, *44*(6), 508–525. https://doi.org/10.1134/S0361768819010031

Fan, Z., Song, X., & Shibasaki, R. (2014). CitySpectrum: A non-negative tensor factorization approach. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 213–233. https://doi.org/10.1145/2632048.2636073

García-Palomares, J. C., Gutiérrez, J., & Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, *35*(1–2), 235–246. https://doi.org/10.1016/j.apgeog.2012.07.002

Hadjidimitriou, N. S., Lippi, M., & Mamei, M. (2021). A Data Driven Approach to Match Demand and Supply for Public Transport Planning. *IEEE Transactions on Intelligent Transportation Systems*, *22*(10), 6384–6394. https://doi.org/10.1109/TITS.2020.2991834

Haghighat, A. K., Ravichandra-Mouli, V., Chakraborty, P., Esfandiari, Y., Arabi, S., & Sharma, A. (2020). Applications of Deep Learning in Intelligent Transportation Systems. *Journal of Big Data Analytics in Transportation*, *2*(2), 115–145. https://doi.org/10.1007/s42421-020-00020-1

Huy Phan, A., & Cichocki, A. (2011). PARAFAC algorithms for large-scale problems. *Neurocomputing*, *74*(11), 1970–1984. https://doi.org/10.1016/j.neucom.2010.06.030

Jung, I., Kim, M., Rhee, S., Lim, S., & Kim, S. (2021). MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.682841

Kim, J., & Park, H. (2012). Fast nonnegative tensor factorization with an active-set-like method. In *High-Performance Scientific Computing: Algorithms and Applications* (Vol. 9781447124375, pp. 311–326). Springer-Verlag London Ltd. https://doi.org/10.1007/978-1-4471-2437-5_16

Liutkus, A., Durrieu, J.-L., Daudet, L., Richard, G., An, G. R., & Richard, G. (2013). *An overview of informed audio source separation*. 1–4. https://doi.org/10.1109/WIAMIS.2013.6616139ï

Mørup, M., Hansen, L. K., & Arnfred, S. M. (n.d.). *Communicated by Terrence Sejnowski Algorithms for Sparse Nonnegative Tucker Decompositions*.

Mützel, C. M., & Scheiner, J. (2022). Investigating spatio-temporal mobility patterns and changes in metro usage under the impact of COVID-19 using Taipei Metro smart card data. *Public Transport*, *14*(2), 343–366. https://doi.org/10.1007/s12469-021-00280-2

Open_Data_Dictionary_CMS_Trip_Survey_2019v2. (n.d.).

Paatero, P., & Tappert, U. (1994). POSITIVE MATRIX FACTORIZATION: A NON-NEGATIVE FACTOR MODEL WITH OPTIMAL UTILIZATION OF ERROR ESTIMATES OF DATA VALUES*. In *ENVIRONMETRICS* (Vol. 5).

Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2015). Comparative framework for activity-travel diary collection systems. *2015 International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2015*, 251–258. https://doi.org/10.1109/MTITS.2015.7223264

Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2017). Transportation mode detection–an in-depth review of applicability and reliability. *Transport Reviews*, *37*(4), 442–464. https://doi.org/10.1080/01441647.2016.1246489

Prelipcean, A. C., Susilo, Y. O., & Gidófalvi, G. (2018). Collecting travel diaries: Current state of the art, best practices, and future research directions. *Transportation Research Procedia*, *32*, 155–166. https://doi.org/10.1016/j.trpro.2018.10.029

Rošťáková, Z., Rosipal, R., Seifpour, S., & Trejo, L. J. (2020). A comparison of non-negative tucker decomposition and parallel factor analysis for identification and measurement of human EEG rhythms. *Measurement Science Review*, *20*(3), 126–138. https://doi.org/10.2478/msr-2020-0015

Shashua, A., & Hazan, T. (n.d.). *Non-Negative Tensor Factorization with Applications to Statistics and Computer Vision*.

Tang, J., Wang, X., Zong, F., & Hu, Z. (2020). Uncovering spatio-temporal travel patterns using a tensor-based model from metro smart card data in Shenzhen, China. *Sustainability (Switzerland)*, *12*(4). https://doi.org/10.3390/su12041475

UC Irvine UC Irvine Electronic Theses and Dissertations Title Understanding the Travel Behaviors and Activity Patterns Using Household-based Travel Diary Data: An Activity Space-based Approach in a Developing Country Context Permalink Publication Date. (2021). https://escholarship.org/uc/item/9w99q3tw

van Wee, B. (2018). Land use policy, travel behavior, and health. In *Integrating Human Health into Urban and Transport Planning: A Framework* (pp. 253–269). Springer International Publishing. https://doi.org/10.1007/978-3-319-74983-9_13

Wang, D., Cai, Z., Cui, Y., & Chen, X. (2022). Nonnegative tensor decomposition for urban mobility analysis and applications with mobile phone data. *Transportmetrica A: Transport Science*, *18*(1), 29–53. https://doi.org/10.1080/23249935.2019.1692961

Wang, Z., Li, X., Zhu, X., Li, J., Wang, F., & Wang, F. (2023). Big data-driven public transportation network: a simulation approach. *Complex and Intelligent Systems*, *9*(3), 2541–2553. https://doi.org/10.1007/s40747-021-00462-2

Welling, M., & Weber, M. (n.d.). *A Constrained EM Algorithm for Independent Component Analysis*.

Zhou, G., Cichocki, A., Zhao, Q., & Xie, S. (2014). Nonnegative matrix and tensor factorizations: An algorithmic perspective. *IEEE Signal Processing Magazine*, *31*(3), 54–65. https://doi.org/10.1109/MSP.2014.2298891

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa