Article

Exploring Hybrid-Multimodal Routing to Improve User Experience in Urban Trips

Diego O. Rodrigues [1,2,*], Guilherme Maia [3], Torsten Braun [2], Antonio A. F. Loureiro [3], Maycon L. M. Peixoto [1,4] and Leandro A. Villas [1]

[1] Institute of Computing, University of Campinas, Campinas 13083-852, Brazil; mayconleo@lrc.ic.unicamp.br or maycon.leone@ufba.br (M.L.M.P.); leandro@ic.unicamp.br (L.A.V.)
[2] Institute of Computer Science, University of Bern, 3012 Bern, Switzerland; torsten.braun@inf.unibe.ch
[3] Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil; jgmm@dcc.ufmg.br (G.M.); loureiro@dcc.ufmg.br (A.A.F.L.)
[4] Department of Computer Science, Federal University of Bahia, Salvador 40170-110, Brazil
[*] Correspondence: diego@lrc.ic.unicamp.br or diego.oliveira@inf.unibe.ch

Abstract: Millions of individuals rely on urban transportation every day to travel inside cities. However, it is not clear how route parameters (e.g., traffic conditions, waiting times) influence users when selecting a particular route option for their trips. These parameters play an important role in route recommendation systems, and most of the currently available applications omit them. This work introduces a new hybrid-multimodal routing algorithm that evaluates different routes that combine different transportation modes. Hybrid-multimodal routes are route options that might consist of more than one transportation mode. The motivation to use different transportation modes is to avoid unpleasant trip segments (e.g., traffic jams, long walks) by switching to another mode. We show that the possibility of planning a trip with different transportation modes can lead to improvement of cost, duration, and quality of experience urban trips. We outline the main research contributions of this work, as (i) an user experience model that considers time, price, active transportation (i.e., non-motorized transport) acceptability, and traffic conditions to evaluate the hybrid routes; and, (ii) a flow clustering technique to identify relevant mobility flows in low-sampled datasets for reducing the data volume and allow the execution of the analytical evaluation. (i) uses a Discrete Choice Analyses framework to model different variables and estimate a value for user experience in the trip. (ii) is a methodology to aggregate mobility flows by using Spatio-temporal Clustering and identify the most relevant of these flows using Curvature Analysis. We evaluate the proposed hybrid-multimodal routing algorithm with data from the Green and Yellow Taxis of New York, Citi Bike NYC data, and other publicly available datasets; and, different APIs, such as Uber and Google Directions. The results reveal that selecting hybrid routes can benefit passengers by saving time or reducing costs, and sometimes both, when compared to routes using a single transportation mode.

Keywords: big data; flow clustering; intelligent transportation systems; multi-source data analyses; spatio-temporal data analyses; user experience

1. Introduction

Intelligent technologies allow industries and governments to serve citizens better through disruptive applications. Among them, we highlight the ones that are related to Intelligent Transportation Systems (ITS), which are responsible for ensuring efficient and sustainable transportation [1]. Many of these applications explore data from several sources to better understand the city dynamics and adapt to them. This constant search for improvement in urban transportation stimulates new ideas and techniques to help city planners and public administrators better understand urban mobility dynamics.

The study of human mobility in urban scenarios has gained more attention recently due to the popularization of location-tracking systems [2], such as Global Positioning Systems (GPS), cellphones, and interactions in location-based social networks. These tracking solutions allow a better understanding of people's movement in a city using buses, subways, taxis, and other means. This understanding is obtained by analyzing their digital footprints. Some smart mobility solutions use data that were collected from human mobility to propose, for instance, urban routing, which focuses on achieving different objectives while suggesting routes to perform trips in the city.

Most of the current urban routing applications focus on the shortest or fastest route identification problem; however, different aspects may play essential roles in route selection. In this sense, several studies have investigated different aspects of urban trips. For instance, routing approaches focused on non-motorized means of transport, also called active transportation (e.g., walking, bicycle) [3], enhancing the levels of enjoyment in the trip [4], or even identifying users' preferences using probabilistic frameworks [5].

This paper investigates the use of different transportation modes as a substitute to traditional modes. We devise a model that allows the evaluation of transportation systems in cities with various transportation modes. The same model can be used to recommend routes by also applying user preferences. We describe both cases in this work. Our solution creates hybrid-multimodal routes that combine different transportation modes on a trip. This approach differs from traditional multimodal routing, which suggests multiple options, each with a single transportation mode. We use historical trip data from taxis of New York City to evaluate the hybrid-multimodal routes. The main objective of this study is to evaluate how a different mechanism for creating and recommending routes for urban trips can impact the user. Our model's main contribution is the recommendation of urban trips that is aware of not only cost and duration, but that also is subjective to users' perceptions. Approaches to model perception of users when recommending urban routes have recently gained relevance in the context of smart cities, for instance, with eco-friendly route recommendation [6] being adopted in mainstream route recommendation apps, such as Google Maps.

This work extends our previous study [7], where we proposed a method to improve route selection through hybrid private vehicles and traffic information. Previously, the route selection was made by evaluating the duration and cost of urban trips. Here, we advance our previous work by exploring a novel user experience framework that is based on Discrete Choice Analysis [8]. This approach allows us to compare different route options while accounting for variables, such as the acceptability of active transportation modes, cost of transfer between different modes, and user preferences through a user profile operator. Besides that, we also implement other improvements, such as consider more alternative modes (e.g., bicycle) and other datasets to evaluate users' impressions of urban trip segments. The novel research contributions of this work are: (i) an origin-destination flow clustering technique that groups urban mobility flows and classifies them in trending or secondary flows; and, (ii) the proposal of a user experience model and a profile operator to evaluate urban trips. Additionally, we evaluate more transportation modes than our initial work and propose a strategy to deal with anonymized data.

The rest of this paper is organized, as follows: Section 2 discusses studies that are related to the proposed technique, i.e., clustering, modeling, and routing. Section 3 provides an overview of the methods used to evaluate our model, which comprises the proposed flow clustering methodology. Section 4 presents our model to evaluate route options for a given trip and how it can be personalized for each user. Section 5 analyzes the results. Finally, Section 6 presents the concluding remarks.

2. Related Work

We study different areas to produce the hybrid multimodal routing methodology. In particular, we propose a mobility flow clustering technique, a hybrid-multimodal routing algorithm, and a user experience model for urban trips. Below, we discuss proposals

for clustering mobility data (Section 2.1); and, routing algorithms and models for route selection (Section 2.2).

2.1. Clustering

The clustering of mobility flows is a research field that creates algorithms to identify trip patterns. There are two main types of clustering in urban scenarios: (i) trajectory-based clustering: the whole trajectory of the moving entity is known, and used [9,10]; and, (ii) Points-of-Presence based clustering: only a few points that are related to the moving entity are known [11–13]. Their main difference is the sampling rate of observed points. In this work, we focus on exploring datasets with data regarding pick-up and drop-off locations only; thus, our proposed algorithm fits into the second type of clustering.

Pang et al. [11] proposed a method to cluster the pick-ups and drop-offs of taxi trips. They spread the points in a fine-grained matrix and write each cell's frequency in it, which is decomposed and factorized. The resulting values are plotted as a heatmap over the city's map. Their method identifies hotspots of trips in the city, which allows them to evaluate their correlation with public transportation stations, business areas, and airports. The structure of the outcomes is the main difference to our solution. Their method focuses only on identifying regions, and ours in identifying the regions and the flow between them, which is an important aspect in the route selection.

Stepwise Spatio-Temporal Flow Clustering (SSTFC) [14] is a spatio-temporal flow clustering technique that identifies mobility trends. SSTFC has spatial and temporal steps and it works likewise other proposed mechanisms in tge literature [15]. In the spatial step, a minimum neighborhood and a size coefficient are used with a custom distance metric to decide whether to merge or not neighboring clusters. Similar parameters exist in DBSCAN [16], namely *min_samples* and *eps*, which are also used to evaluate the density of a region and create clusters. After the spatial step, SSTFC uses the same method to create temporal clusters inside the spatial ones, but using a third parameter, a temporal threshold, to merge neighboring clusters. In [14], it is assumed that SSTFC is robust, but there is no quantified assessment. As clusters get merged several times, flows with varied directions are mixed in the same group, which is sub-optimal when trying to identify mobility trends. Our algorithm first locates the functional departure and arrival zones, which limits the flow directions. These flows can have different angles, but they always flow from one region to another, while, in their technique, flows going forward/backward might end at the same cluster.

2.2. Routing and Modeling

Different factors affect decision making when choosing urban trip routes. Most commercial systems and IT studies map the problem to the shortest/fastest pathfinding [5,17]. The literature of Discrete Choice Modeling [8] and Transport Economics [18] for route choice modeling is well-studied. Yet, some proposed models are prohibitive due to data restrictions that are present in real-world applications and the difficulty in assessing every aspect that drives route selection. Our model builds upon discrete choice modeling frameworks and defines premises to ease the implementation in the real world. Additionally, we identify data sources that can be used to better model the user's behavior.

This study assumes that cost, duration, and quality of experience (QoE) play a central role in the mode/route choice process. Other studies model these choices while considering more variables. The FAVOUR algorithm [5], for instance, uses a probabilistic model over historical data. It represents the route's cost/utility as a weighted sum of the cost of route segments. Every segment is evaluated with different features, leading to different weights for the same segment. Furthermore, the algorithm considers that some route options have costs that are associated with the entire route, rather than every segment to contemplate different phenomena, such as weather. We use similar variables to those that are used in FAVOUR, such as distance, duration, price, and the number of transport mode changes. However, the FAVOUR algorithm does not consider the acceptability of

using active transportation modes, which we do. In their study, walking and cycling long distances are evaluated in the same way as short ones. Another difference is that we combine the features into a familiar metric to users, such as a monetary price or a duration. The FAVOUR model is based on a Bayesian learning technique to define the weights for the sum and recommend routes for users. Differently, we use Discrete Choice Analysis [8] to model the costs of a trip and apply nominal values for the quality of service (QoS). Finally, we use a profile operator to suggest routes separately that provide ways to compare routes with or without the bias of individuals. FAVOUR always counts on the existence of individual user data to adapt its recommendations. Finally, our model considers traffic conditions to select transition mode points, while FAVOUR does not consider real-time traffic information.

Hrnčíř et al. [19] proposed a method for selecting bicycle routes using different criteria. They defined a generic framework based on a graph with geo-coordinates, altitude, the horizontal length of cycling paths segments, and other features of route segments. For the experiment, they defined the features to be the ones that are provided by OpenStreetMap (https://www.openstreetmap.org/, accessed on 14 May 2021). The authors modeled the route selection problem as multi-criteria search optimization and proposed heuristics to make this search feasible. This study focuses on maximizing different utility functions that are applied to routes. They do not provide methods to obtain scores that aid in ranking or personalizing the results, such as our proposal. They only consider one transportation mode (i.e., bicycles), which causes important variables to be omitted, such as the acceptability to chose a given route using active transportation. When taking active transportation, users may feel discouraged to perform long trips, which is considered in our model.

3. Evaluation Framework

This work proposes a method to combine different transportation modes for single urban trips. For each trip, we evaluate its duration, price, and user experience. Section 3.1 briefly introduces the concept that we use to evaluate urban trips. Section 3.2 describes the methodology used to combine transportation modes, produce multimodal-hybrid routes, and its evaluation. This solution explores real-time traffic data to replace trip segments using taxis with other transportation modes (e.g., bicycle or public transportation). In order to evaluate the hybrid-multimodal routes, we use datasets from the city of New York, available at the NYC Open Data portal (https://opendata.cityofnewyork.us/, accessed on 14 May 2021). We performed data reduction by clustering the main mobility flows due to their size, as described in Section 3.3. When looking for route choices, we run different nearest neighborhood and temporal queries. To reduce time executing these queries, we combine two different data indexes into a single data structure creating the Multi-Layer Geographical Linkstream, as detailed in Section 3.4.

3.1. Overview

Figure 1 illustrates the process of evaluation of urban trips. We start with a dataset containing origins and destinations of urban trips. After, we use our proposed mobility flow clustering technique, as described in Section 3.3, to create a weighted graph of the urban flows. This step makes feasible the evaluation of huge datasets by applying the model on the most relevant flows. Once the graph is obtained, we evaluate possible driving ways for a given flow and combine these driving ways with traffic data to identify congested segments in the trip. These segments then are replaced with alternative transportation modes in the route combination phase. After all traditional and hybrid route options are obtained, we assess the gains/losses in terms of duration and cost, and also we use our proposed user experience model (as described in Section 4) to evaluate the quality of all options.
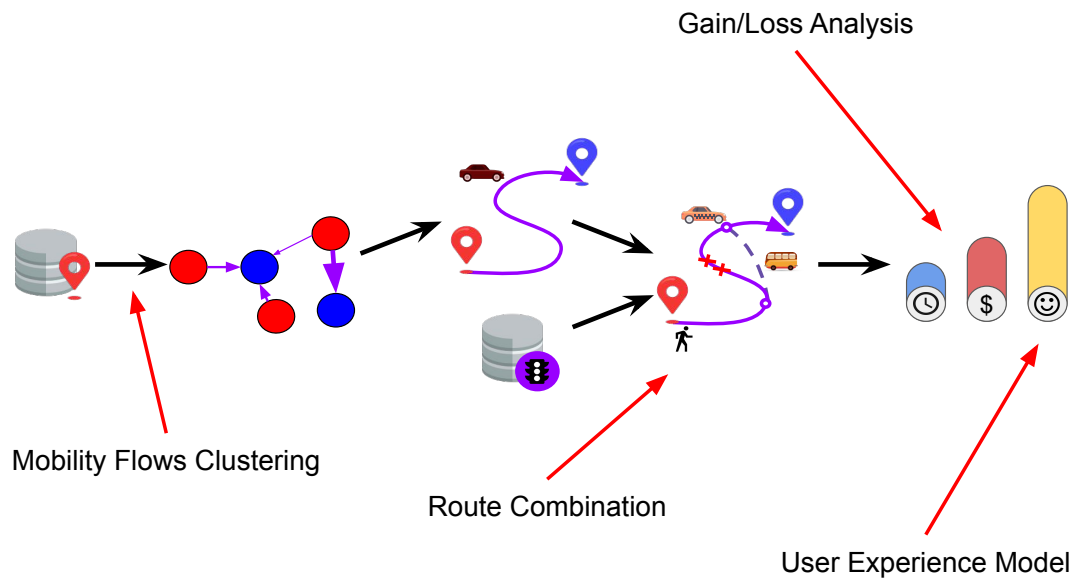
Figure 1. Overview of proposed methodology.

3.2. Methodology

We propose a Multi-Layer Geographical Linkstream (MLGLS) to organize data from different sources to assess the impacts of using multimodal hybrid routes for urban trips. We study the impacts of replacing taxi-only trips in New York with hybrid options using our proposed methodology, such as reducing time spent on traffic jams, which leads to an increase in QoE during the trip. These hybrid routes contrast to the traditional approach of multimodal routing, which consists of different options using single modes (e.g., one public transportation, one driving, and one walking option). The hybrid approach that was proposed in this work also considers routes that started with one of the traditional modes and finished in another (e.g., started in public transportation and finished in a taxi). The purpose of this section is to describe the methodology used to evaluate the hybrid routes. The process is divided into seven steps: (i) reduce data about trips from Yellow and Green Taxis datasets and find main mobility flows; (ii) create and populate an MLGLS with trip data and with other transportation modes (e.g., bus stations, bike dockers); (iii) evaluate the initial driving options that contain traffic information; (iv) identify congested segments in the initial options; (v) select mode transition points and combine them; (vi) trace route alternatives and save them; and, (vii) evaluate the metrics of the alternatives. Steps i, ii, and vii are executed once for the whole dataset. In contrast, the other steps are repeated to every mobility flow in Step i. For each one of these flows, the trip is evaluated when considering the spatio-temporal ⟨origin, destination⟩ pair. Additionally, changes in the scenario during the trip are not accounted for. Figure 2 outlines the methodology.
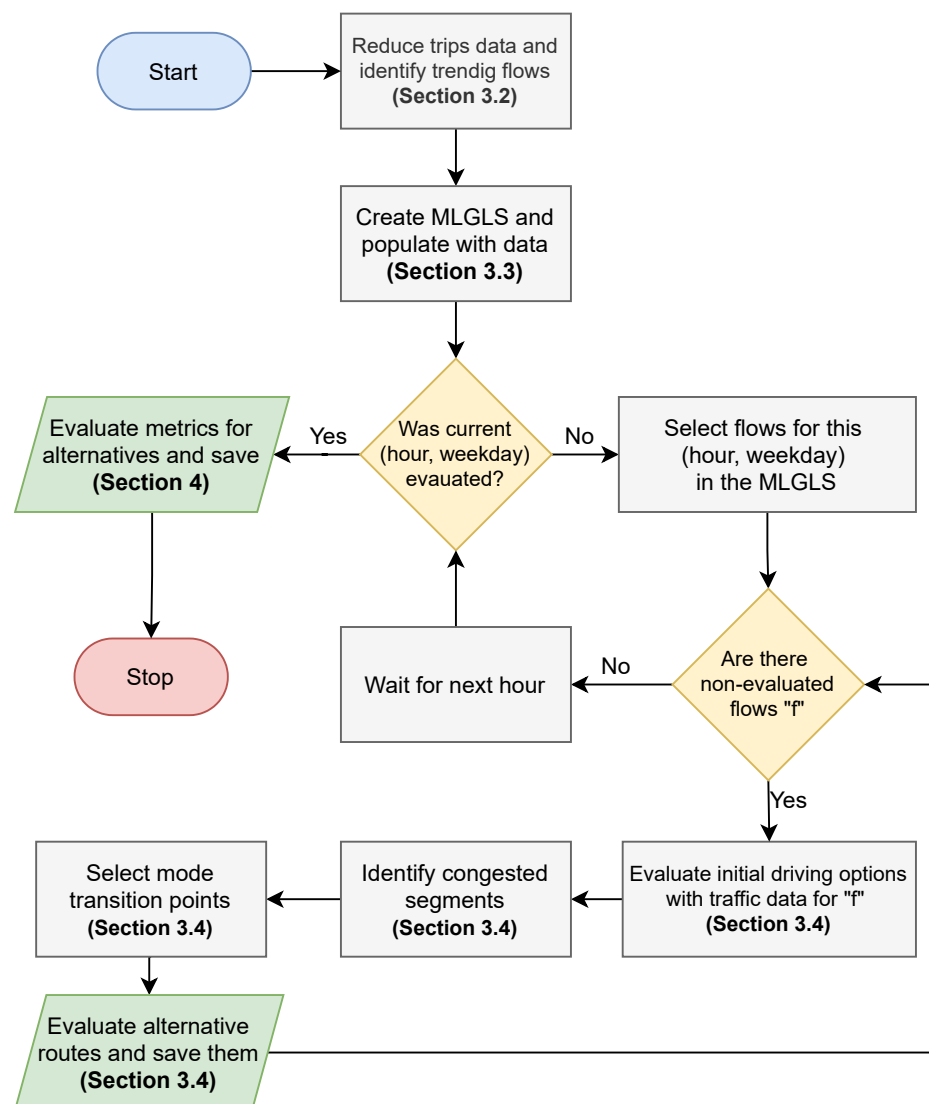
Figure 2. Work flow to evaluate the impact of using hybrid-multimodal routes to replace taxi trips in New York.

Section 3.3 describes Step i, Section 3.4 give details about the MLGLS used in Step ii. Section 3.5 comprises Steps iii. iv, v, and vi. Section 5 explains the outcomes that are produced in Step vii, and other important intermediary results, such as the individual route selection evaluation using the model described in Section 4.

3.3. Data Reduction

The proposed model analyzes trip-by-trip, every route option by requesting traffic data, trip fares, and route duration estimates through different APIs in real-time over the internet to assess the price, duration, lengths, and waiting times of trip segments. This analytic approach demands a long time to perform data collection and analysis. Thus, to have a feasible study case, we have a data reduction step that clusters mobility flows. This step aims to identify the city's main flows to perform the evaluation. The data reduction technique that is described below is general enough to be used in different datasets and for different purposes.

The technique identifies trip Origin-Destination Mobility Flow Corridors (OD-MFC). This is accomplished by finding functional zones on the coordinates of the dataset and then computing the flows between these zones. It has two variants to adapt to different aspects of datasets. If the dataset has geographical coordinates (i.e., latitude and longitude), the

full version of the technique must be applied. However, there are plenty of datasets that are anonymized, only containing region ids where the departures and arrivals happened. For these datasets, a light version of the technique was developed.

Both of the versions of the OD-MFC technique have three steps: (i) Split (pre-processing): divides the dataset into chunks by considering a division between hours, weekdays, date or other variants—we used the division ⟨hour, weekday/weekend⟩, e.g., ⟨19 h, weekend⟩, which describes the class with trips between 19 h and 20 h on Saturday and Sunday; (ii) Functional Region Identification: clusters the departures and arrivals of trips to identify the most relevant zones; (iii) Flow Accounting: is responsible for accounting the flows between those identified zones; and, (iv) Flow Classification: classifies the flows in trending (i.e., flows with a more relevant amount of trips) and secondary.

The pre-processing step divides the dataset into smaller ones that capture the data seasonality. In our study, we defined this division by analyzing different shapes of distribution curves of the trips. Afterwards, the second step identifies the functional regions in the data (i.e., the most relevant zones). This is the single step in which the full and the light versions of the technique differ. For the full version, we use the HDBSCAN [20] clustering algorithm to identify the functional zones. This algorithm is widely adopted to cluster positional datasets containing noise. For the light version, the regions are already identified. Thus, all of the regions could be used in the analysis.

The third step consists of accounting for the flows between regions. Each flow is described by the number of trips between two regions. Finally, the fourth step identifies the main flows. The classification technique was inspired by a similar strategy that was proposed to identify the EPS in the DBSCAN algorithm [16]. In their paper, the authors evaluate the k-nearest neighbor distance for all samples in the dataset and plot them sorted. This plot tends to form an exponential decay curve. Based on this technique, we created a that sorts the flows based on their magnitude and identifies the "knee" of the resulting curve. We use Equation (1), the Curvature for a Plane Curve [21] equation, to identify this "knee".

$$\kappa = \frac{|y''|}{(1 + y'^2)^{3/2}}. \tag{1}$$

The function y is the curve that formed by the sorted flow magnitudes. This equation evaluates the magnitude of the curvature at a certain point in the curve. Because we are looking for the "knee" of the curvature, we must look for local maximum for κ, i.e., $\kappa' = 0$. A previously faced problem of this study [7] was that the curvature formed by real data was not a smooth curve. Thus, to enhance calculating the required derivatives, we now smooth the data, depending on the desired results, before applying the classifier. In this work, we use interpolation with an exponential function to smooth our data. We select the "knee" point in the curve, which is used as a threshold to classify the flows in trending or secondary. If we use an exponential function of the form $y = e^{ax}$ to model the exponential decay, then the solution of Equation (1) leads to:

$$\kappa = \frac{a^2 e^{ax}}{(1 + a^2 e^{2ax})^{3/2}}, \tag{2}$$

where the maximum is given at:

$$\kappa' = 0 \longrightarrow x = -\frac{\log 2a^2}{2a}. \tag{3}$$

In this study, we use HDBSCAN algorithm [20] to identify functional regions. HDBSCAN is a $O(n^2)$ time algorithm, with approximate solutions being given in average $n \log n$ time [22]. The proposed classification approach has to first sort mobility flow popularity in time $O(m \log m)$, with $m \leq n$, but, in general, $m << n$. After that, the approach needs to solve a minimization problem to identify the best fit exponential equation to fit the

curve that formed by the sorted popularities. We use the Levenberg-Marquardt method, which takes $O(\epsilon^{-2})$ time [23]—ϵ is the precision adopted. After identifying the best fit curve, the classification is done in $O(1)$. Therefore, the approach is executable in $O(p^2)$, where $p = \max\{n, 1/\epsilon\}$. The proposed method maintains the same time complexity of HDBSCAN.

To assure the quality of the proposed clustering technique, we first verify whether it complies with the requirements of good clustering techniques [24], i.e., the clustering: (i) should not impose a priori bias on the clusters' shape; (ii) should be able to handle varying densities; and, (iii) should be able to handle varying dimensionality. The origin and destination functional regions are obtained using the HDBSCAN technique, which does not impose a bias on the clusters' shape, and it is also known for good performance with noisy datasets. Additionally, the selection of the trend/secondary flow threshold identifies a point with no bias, thus meeting requirements (i) and (ii) (as discussed in Section 5.3). Regarding requirement (iii), despite using a spatial 2D dataset in this study, neither HDBSCAN nor the curvature equation limits the number of dimensions of the dataset. Thus, the proposed technique can handle higher dimensionalities. We conduced a experiment comparing our approach to the SSTFC algorithm [14]. Section 5 shows that both versions of OD-MFC outperform SSTFC when identifying the most relevant flows in noisy datasets.

3.4. Data Structure

We model all the data in MLGLS data structure, which was created by combining: (i) Linkstreams [25], a temporal graph structure that stores links sorted according to time of occurrence; (ii) KDTree [26], which is a tree structure to ease nearest neighbor searches; and, (iii) Multi-Aspect Graph [27], a derivation of static graphs that allow the creation of multiple layers to represent data features.

A link is a 3-tuple (a, b, t) that shows a contact between a and b at time t. Contacts that are represented in this stream are time-sorted and they may optionally have a duration, in which case the Linkstream may be called a Stream Graph. Because links are sorted in time, one can use a binary search to optimize queries. However, the Linkstreams are not designed to deal with positions, an important feature of mobility data. To tackle this issue, we added a complementary index to the Linkstream using KDTrees. The way that this tree is constructed allows the quick execution of neighborhood searches, as opposed to other spatial trees, such as the R*-tree [28].

The use of Linkstream and KDTree makes it possible to run quick spatio-temporal data searches. However, urban mobility is heterogeneous and it can be collected from different sources, and must be combined to perform a more insightful analysis [29]. Such a process must not lose information about the data origin since data from different contexts may have different interpretations. To combine data without losing its source information, we use a feature from the Multi-Aspect Graphs, the possibility of creating multiple layers in the graph to represent different data sources. This data structure allows for the creation of aspects, a data dimension where multiple layers can be added. Hence, we propose the usage of several layers where data from different sources can be sheltered.

Formally, the MLGLS is a group of nodes $v \in V$, connected by links $e \in E$. These nodes and links may hold further information detailing the event identified by the node or the type of contact that is indicated by the edge. The existence of events (or nodes) are known via a data source (or layer) $l \in L$; and, all of the interactions happen in a time interval $T = [t_0, t_f]$. Additionally, to enhance the spatial queries' performance, we have a Geographical Index I (i.e., the KDTree). Thus, the MLGLS is given by the tuple $M = \langle V, E, L, T, I \rangle$.

An important implementation aspect is the existence of different layer types. The data structure supports both static graphs and linkstream layers. For static layers, the difference is that nodes and links exist during the whole time interval, from t_0 to t_f. Because of the characteristics of modern datasets, there is another type of layer, called dynamic. This type

is used for layers in which nodes and edges vary in time, but the information regarding these graph entities can only be obtained when needed. For example, the bike dockers in our experiments may or may not have bicycles available at a given moment. Thus, when we need to verify the existence of a node, we perform a call to the provider API to evaluate if that node will, or will not, be listed in the graph. The dynamic layer is defined in the MLGLS as a time-dependent function called upon the necessity to evaluate the current state of a layer.

Figure 3 depicts our proposed data structure, which has two dimensions: time and data sources. Layers can be created at each dimension, being represented by rows or columns of rectangles in Figure 3. The dotted blue line L_h is surrounding the taxi data layer in the data-source dimension, and the green dotted line L_v is surrounding the time layer t_2. The nodes (circles) are the data entries, and each link (continuous line) represents the contact between those entries. The dashed thick lines represent contacts that persisted in time. Finally, the red lines represent the spatial index that is created by the KDTrees. This index creates perpendicular divisions in the space to separate nodes and ease queries.
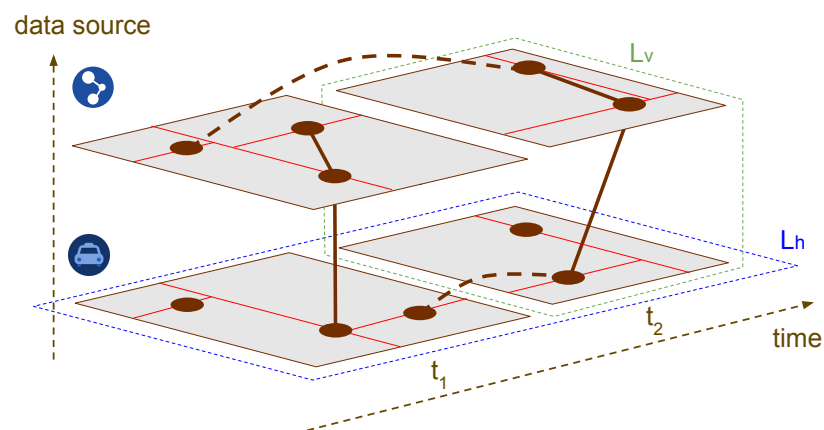

Figure 3. Example of Multi-Layer Geographical Linkstream.

When merging transportation modes to create alternative urban routes, we fuse five positional datasets: subway stations, bus stops, bicycle dockers, and the Yellow and Green Taxis trips. When analyzing these datasets, we have to query data in time and space; the proposed data structure saves time and resources. Additionally, it is general enough for usage in other scenarios. The approach used to combine these datasets and produce hybrid-multimodal route alternatives is described below.

3.5. Route Combination

The processes between Steps iii and vi (Section 3.2) are repeated for every flow inside its respective ⟨hour, weekday⟩ class. A flow is defined by an origin, a destination, and a weight based on the number of trips. For every flow, we obtain a list of alternative routes, including the traditional, and the hybrid options to be compared. Algorithm 1 [7] shows how this list of alternatives is produced.

Algorithm 1 Evaluate route alternatives for trip flows.

Input: Origin and destination of a flow.
Output: List with the route alternatives for the flow.

1: *drive_way* ← get_driving_way(*origin*, *destination*)
2: *start_transitions* ← *new List*()
3: *end_transitions* ← *new List*()
4: *options* ← *new List*()
5: *previous_step* ← **none**
6:
7: push(*start_transitions*, *origin*)
8: push(*end_transitions*, *destination*)
9:
10: **for** *step* **in** *drive_way.steps* **do**
11:    **if** is_congested(*step*) **then**
12:       **if** is_congested(*previous_step*) **then**
13:          pop(*end_transitions*)
14:          push(*end_transitions*, *step.destination*)
15:       **else**
16:          push(*start_transitions*, *step.origin*)
17:          push(*end_transitions*, *step.destination*)
18:       **end if**
19:       *previous_step* ← *step*
20:    **end if**
21: **end for**
22:
23: **for** *st* **in** *start_transitions* **do**
24:    **for** *et* **in** *end_transitions* **do**
25:       *opts* ← get_options(*origin*, *st*, *et*, *destination*)
26:       concat(*options*, *opts*)
27:    **end for**
28: **end for**

Algorithm 1 starts (Line 1) by calling the `get_driving_way` function (Step iii), with the origin and destination as arguments. This function calls the TomTom Routing API (https://developer.tomtom.com/, accessed on 14 May 2021), which returns a list of steps to perform the trip and traffic data. In Lines 7 and 8, the algorithm appends the origin and destination to the list of transition starts and ends. They are added to create the possibility of full trips being performed in single modes (i.e., the traditional routes). The `foreach` command in Line 10 populates the transition lists according to the state of traffic (Step iv). This state is evaluated via the function `is_congested`, which indicates whether it has traffic or not. This binary classification is only used to produce the route options. When evaluating the Quality of Experience (QoE), we consider the time that would be spent on the congestion, which is assessed based on historical data using the TomTom API. Note that the steps with bad traffic conditions in a row are merged in the **if** command in Line 12 just by updating the `end_transitions` list.

Lines 23 and 24 contain the `foreaches` that are responsible for combining the transition points (Step v). Inside these `foreaches`, a call to the `get_options` function evaluates all alternatives considering a tuple with an origin, destination, start transition, and end transition. This function uses the data populated in the MLGLS to identify the actual transition points, when considering mode characteristics (e.g., bus stops). Next, the function evaluates the possible options and saves data about traffic, waiting times, duration, and prices for every step and the mode used to compose it.

The resulting list of options is concatenated to a list with all options (Line 26) to be saved, as described in Step vi. After evaluating all of the route alternatives, they are compared with the traditional routes using the metrics duration, price, and user

experience (Step vii). Section 4 describes how to evaluate the user-experience model. The route combination and analysis code is open source, being distributed under the SMAFramework [30] on Github (https://github.com/diegopso/smaframework, accessed on 12 May 2021).

4. User Experience Model

This section describes the evaluation of user experience in urban routes, which is based on the Opportunity Value [31]. This value is a technique used in Discrete Choice Analysis to assess the utility of a decision made in a set of mutually exclusive options. This decision is the same when picking a route in an urban trip: once a route is chosen, the others become unavailable. Sections 4.1–4.3 model the different deterministic aspects of the route. The models in these sections can be used to evaluate the transportation system conditions and various transportation modes. Section 4.4 presents the experiment instances that are used to evaluate user satisfaction. Section 4.5 introduces a Profile Operator used to quantify the impact of personal choices in the model. It works on top of the deterministic evaluation and adds a probabilistic dimension regarding individual preferences and, thus, enables the model to be used when performing route recommendations.

4.1. Core

Let $r \in R$ be a route option to a given trip of a person (decision-maker). R is the set of available options created with a heuristic. r is a tuple $\langle t_r, c_r, l_r, S_r \rangle$, the route's duration, cost, length, and set segments in different transport modes, respectively. One way of comparing aspects of route options of a trip is evaluating the opportunity cost of these aspects. For instance, the Opportunity Cost Equation [31] (Equation (4)) evaluates the opportunity cost for selecting a route in the set R as the loss due to selecting this option and not the best possible option.

$$copp_t(r) = t_r - \min_{\forall r' \in R} t_{r'}. \tag{4}$$

where r and r' are the routes taken from the set R with duration t_r. $copp_t(r)$ denotes the opportunity cost of duration of a given route r. The opportunity costs of other route aspects are given by analog equations evaluated on c_r, l_r, and so on.

A typical use of the opportunity cost in transport engineering is to evaluate the Value of Time Travel Savings (VTTS). It represents the monetary cost of the extra time that is spent by selecting a given route instead of the fastest route.

Nevertheless, in order to obtain the user experience in the trip, we must consider both the value of the user's time and the actual paid value, instead of only the duration of t_r. We can use the Generalized Cost Equation [32] (Equation (5)), also used in transport economics when evaluating the attractiveness of a route option [33], to combine the two variables. The generalized cost of a route option is the addition of the paid monetary cost to the generalized journey time, which was found via a utility function.

$$g(r) = c_r + u(r). \tag{5}$$

where $g(r)$ is the generalized cost, c_r the monetary cost of the route, and $u(r)$ the utility function to estimate a generalized time. The generalized time u is estimated using the Value of Time Travel (VTT) of the user. u varies with the context of the decision-maker, such as route conditions and his income/hour. One way of evaluating u is by multiplying the trip duration by the VTT of the decision-maker, as $u(r) = \tau t_r$, where τ is given in [USD/h]. This may be used to compare the routes with similar characteristics. However, when comparing multiple transportation modes, the impressions of the decision-maker might vary among different modes. One way of contemplating this phenomenon is by using a Generalized Time Function [34] that considers the user perceptions to adapt the measured time. For instance:

$$u(r) = \tau T(r). \tag{6}$$

Subsequently, to define the general time of the complete route $T(r)$, we sum the general time of all segments $s \in S_r$:

$$T(r) = \sum_{\forall s \in S_r} T(s), \tag{7}$$

where $T(r)$ is the generalized time in the complete route, and $s \in S_r$ is a route segment in a specific transportation mode, which has a duration, cost, length, and its mode $s = \langle t_s, c_s, l_s, m_s \rangle$.

4.2. Mode Transfer Cost

When a user changes transportation mode, he incurs different costs. For instance, when changing to a private hired vehicle, companies usually practice a hiring fee that increases the monetary cost of the trip. Furthermore, some transportation modes require a waiting time, until a vehicle arrives to take the user, which increases the total duration of the trip. Hiring fees and waiting time for private hired vehicles and busses are both assessed in our experiments; they are obtained from real data through different APIs, e.g., Uber and Google Directions. Besides these costs, we use a mode transfer penalty to model the direct costs in terms of experience that the user incur by having to change the transportation mode. We introduce an exponential depreciation operator in Equation (8) to account for mode transfer penalty cost in experience for the user. We count the number n_r of changes in the route and use it to compute the total impact in a given route option as:

$$\lambda_r = \begin{cases} 0, & \text{if } n_r = 0 \\ e^{a n_r}/100, & \text{otherwise} \end{cases}. \tag{8}$$

where λ_r is the impact due to the number of mode changes. In our study, we did not obtain supporting data to evaluate a, this parameter could vary according to user preferences. Yet, $a = 1$ results in a reasonable impact considering traditional urban trips standards. Table 1 shows the impact according to the number of mode changes for $a = 1$.

Table 1. Transfer mode impact for $a = 1$.

n_r	0	1	2	3	4	5
λ_r [%]	0%	2.72%	7.39%	20.09%	54.60%	148.41%

Using Equation (8), the final generalized time of a route in Equation (7) changes to Equation (9).

$$T(r) = (1 + \lambda_r) \sum_{\forall s \in S_r} T(s). \tag{9}$$

4.3. Experience in Different Route Segments

The satisfaction of the decision-maker in a trip segment varies according to segment accessibility. Thus, to measure this satisfaction, the general time for a given segment is:

$$T(s) = S(s)t_s + \epsilon, \tag{10}$$

where $S(s)$ measures user satisfaction in a given transportation mode. The variable ϵ describes the error due to non-observable components, such as weather conditions. Here, we consider that this error is normally distributed, which causes a reduced impact on the final values according to the Random Utility Theory [18].

To define S, we use In-Vehicle Time (IVT) multipliers that estimate impressions of a user in a given transport mode as compared to in-vehicle impressions. For example, a given user may consider 5 min of walking to be as unpleasant as 10 min in a vehicle. In

this case, the IVT multiplier would be $10/5 = 2$. There are different ways of estimating IVT values, which usually consider surveying passengers. We consider the characteristics of the segment s and impressions of the level of service of the decision-maker w to measure the satisfaction S.

$$S(s) = w(s)P(s). \tag{11}$$

$P(s) \in [0,1]$ is a depreciation factor to describe the acceptability of traveling a segment s. Functions P and w depend on the scenario, and their selection is discussed in Section 4.4.

4.4. Experiment Instances

To evaluate the satisfaction of users in route segments in the present study, we consider that both functions w and P vary with the transportation mode. Additionally, the impression about the level of service in a segment depends on the time spent; and, the acceptability of a given segment is a function that depends on the length of the segment, i.e.,

$$S(s) = w_{m_s}(t_s)P_{m_s}(l_s). \tag{12}$$

where m_s is the mode used on the segment s; w_{m_s} and P_{m_s} are respectively the impression of the user and the depreciation function in relation to the transport mode used in that segment. We consider a limited set of modes and use data from a survey [35] to obtain Equation (13), an instance of w. We combine bicycle and walking modes, since both require physical effort to perform, i.e., active transportation. We do not have the actual multiplier for the bicycle mode, but a study conducted in Beijing suggests bike satisfaction to be high when compared to vehicles [36].

$$w_{m_s}(t_s) = \begin{cases} 1.54t_s, & \text{if } m_s \text{ is in-vehicle congestion} \\ 0.78t_s, & \text{if } m_s \text{ is in-vehicle headway} \\ 1.70t_s, & \text{if } m_s \text{ is waiting} \\ 1.65t_s, & \text{if } m_s \text{ is walking or bicycle} \end{cases}. \tag{13}$$

where the impressions of the user w_{m_s} in relation to the transport mode m_s is obtained in function of the duration of the trip segment t_s.

$P_{m_s}(l_s)$ presented in Equation (11) is a depreciation factor that decreases the user satisfaction in a segment according to its length, which depends on the transportation mode m_s. In the urban context, vehicles and public transportation usually perform the longest journeys (modes as helicopters are omitted). Thus, the impact of $P_{m_s}(l_s)$ in such routes is minimal. However, the impact on the walk and bicycle modes is highly relevant and they are usually mentioned as walkability [37] and bikeability [38], respectively. They may consider more factors other than the distance (e.g., accessibility, air pollution).

One way of guessing the likelihood of someone adopting a segment of active transportation is to observe a dataset of trips with their lengths. We can estimate the probability of choosing a segment by calculating an inverse cumulative distribution function of trip lengths in the dataset. This approach is already used in the literature to estimate walking and bicycling acceptabilities (The Netherlands [39] and US [40,41]). Moreover, these abilities vary according to the trip purpose. For everyday commute trips using public transportation, they also depend on the accessed mode—one might walk twice the distance for using rapid modes as metro or tram [41].

In this work, we estimate bikeability using data from the Citi Bike NYC (https://www.citibikenyc.com/, accessed on 14 May 2021) in New York. However, we did not have access to a dataset to estimate walkability. Thus, we use previous results that were obtained for US cities [40]. We defined both functions by interpolating the results for walkability and bikeability. P_w (Equation (14)) is used to evaluate the acceptability when walking, and P_b (Equation (15)) when using a bicycle. We also account for users' will to walk further for accessing rapid transportation modes [41]. The value of l_s for the equations is given in kilometers. Additionally, the values of P are capped between 1 and 0.01 to prevent invalid

outcomes. Section 5.5 discusses walkability and bikeability, and shows the shape of the inverse cumulative distributions.

$$P_w(x) = -6.829x^4 + 9.788x^3 - 2.847x^2 - 1.858x + 1, \tag{14}$$

$$\begin{aligned} P_b(x) = -2.11 \times 10^{-4}x^5 + 3.93 \times 10^{-3}x^4 \\ -2.63 \times 10^{-2}x^3 + 7.50 \times 10^{-2}x^2 \\ -8.05 \times 10^{-2}x + 1.01. \end{aligned} \tag{15}$$

4.5. Individual Route Selection

This section proposes a strategy for selecting routes among different options based on scores and they can be used to create a recommendation set. The recommendation should also consider other issues, such as the impact of recommended routes in the overall system when the adoption rate is high. The recommendation approach is responsible for mitigating this issue by, for instance, load balancing the set of reasonable routes [42]. We do not address the issue of load balancing recommendations in the present study.

The described model has three main results: (i) the total opportunity cost (Equation (4)) to be used to select a route among different choices; (ii) the generalized cost (Equation (5)), which measures an overall cost/benefit of an option considering its cost, users' perception and duration; and, (iii) the generalized time (Equation (9)), which removes the cost and only evaluates the users' perceptions of time. These metrics model general users in specific populations. However, when it comes to individual choices, these values may vary.

We introduce a profile operator to tackle the issue of different opinions of users when selecting routes. This operator is inspired by the concept of Mixed Strategies from Game Theory, which is a probability distribution over a set of actions taken by an agent [43]. We consider that the decision-maker has two Pure Strategies: always choose the shortest duration or always choose the cheapest trip.

To produce a Mixed Strategy, we allow the decision-maker to pick a probability p in which the QoE is favored, despite the cost. The remaining odds $(1 - p)$ lie in the event of choosing a trip based on the cost. The Expected Utility [44] (Equation (16)) of the outcome of selecting routes, based on the Mixed Strategy that was adopted by the decision-maker, is obtained in this case, for two variables, as:

$$\mathbb{E}[P] = u(A)p + u(B)(1 - p), \tag{16}$$

where p is the probability that defines the profile P and $u(A)$ and $u(B)$ are the utilities of the choices A and B, respectively. The difference in our operator is that, in our equation, we use two utility functions for the same action, resulting in Equation (17).

$$\zeta = u_1(r)p + u_2(r)(1 - p), \tag{17}$$

where ζ denotes our proposed operator, $u_1(r)$ and $u_2(r)$ are the two different utility functions for the action of selecting the route r, and p, $(1 - p)$ the respective complementary probabilities. If we select $u_1 = (t + T)/2$, and $u_2 = c$ from the utility functions previously described, then the profile operator ζ of a route r given the profile p becomes, as shown in Equation (18). u_1, represents the QoE-related aspects and u_2 the monetary costs.

$$\zeta(r, p) = p\left[\frac{t_r + T(r)}{2}\right] + (1 - p)c_r. \tag{18}$$

where the profile operator ζ for a given route r and profile p is given in terms of the duration t_r, the generalized time $T(r)$, and the cost c_r, and its related probabilities p and $1 - p$.

In Equation (18), $p \in [0, 1]$ represents the value that a specific user gives to QoE-related variables. QoE represents the duration and overall user impressions about the duration.

We oppose the cost to the quality metrics, i.e., users expecting better quality might pay more for the trip, while a user expecting to pay less may have a worse service. The best route, according to the model, obtains the smallest value for ζ. To apply Equation (18), t_r, c_r and $T(r)$ must be normalized. We use the z-score normalization function for the experiments in this work. We selected the z-score, since it assesses the distance between a route aspect measurement and the average of all measurements of that same aspect for a trip. This evaluation of distance is given in terms of standard deviations; it counts how many times an option is worse than others in terms of standard deviations.

With this operator, our model is complete and it complies with Discrete Choice Modeling framework [18], which is comprised of four elements: (i) the decision-maker: the person that aims to choose a route; (ii) the alternatives: the routes; (iii) the attributes: the utility functions; and, (iv) the decision rule: the profile operator.

5. Results and Discussion

In this section, we discuss the obtained results. We start describing the different data sources used to evaluate hybrid trips in the city of New York in Section 5.1. Afterwards, we discuss the mobility flow clustering technique proposed; the pre-processing results are listed in Section 5.2, the final output of the clustering is presented in Section 5.3, and, in Section 5.4, we compare our proposal with a study from the literature. Finally, we present the results that are related to the hybrid routes and the user experience model. Section 5.5 discusses the acceptability of using active transportation modes, Sectoion 5.6 discusses the overall values for cost, duration, and user experience for the evaluated trips, and Section 5.7 shows the results for the personalizing routes based on user profiles.

5.1. Data Characterization

In this section, we briefly explain the datasets that are used to obtain the results in this study.

5.1.1. Taxi Data

The main datasets used in this work are the Yellow and Green Taxis datasets from New York (https://www1.nyc.gov/, accessed on 14 May 2021), which were taken from the NYC Open Data portal, from 2016 and 2017. Part of the period is anonymized, i.e., they do not contain geographical coordinates of the trips' pickups and drop-offs, but, instead, the IDs of regions where they started/ended. These are the two main taxi services in New York: the Yellow Taxis cover mainly the center of Manhattan Island and airports, and the Green Taxis usually cover the periphery of the city. The Yellow and Green Taxis datasets have 113.4 and 11.7 million trips, respectively. We removed invalid data from these datasets, such as trips starting and ending at the same point, vehicles traveling over 100 km/h, or trips over 6 h. The remaining trips were clustered into flows and evaluated.

5.1.2. Traffic Data

We used TomTom Routes API (https://developer.tomtom.com/, accessed on 14 May 2021) to assess routes and know their traffic conditions. The requests happened at the same hour when the trips occurred to have similar traffic conditions to those when the flow happened.

5.1.3. Bicycle Data

This data are from the Citi Bike service in New York (https://www.citibikenyc.com/, accessed on 14 May 2021), 2017. This service is the most important bicycle sharing system in the city, and we used 257,498 trips from this dataset. We also used docker's location and availability to calculate route alternatives using shared bicycles. Finally, we used historical trip data to evaluate the bikeability.

5.1.4. Routing Data

We used Google Directions API (https://developers.google.com/, accessed on 14 May 2021) and Uber API (https://developer.uber.com/, accessed on 14 May 2021) to guess public transportation and private hired vehicles routes. The Google Directions API gives the available route segments in public transportation and information regarding bus schedules and live positions (when available) to evaluate waiting times. Uber API is used to obtain fare estimations and waiting times for trips.

5.2. Pre-Processing

Before starting the evaluation, the first step of the analysis consists of splitting the dataset into classes for evaluation. Observing the dataset of Yellow Taxis, we found that an hourly division would be enough to capture its seasonality (see Figure 4). It is possible to observe the variation in the number of trips and passengers according to the hour of the day. Additionally, it is possible to visualize the difference in the curve behavior for weekdays and weekends. Figure 4 shows some relevant areas in the curve, high and low peaks of trips on weekdays, and a peak at the weekend. We made an hourly division to create classes, but the route evaluation used real-time data. These relevant areas were used to evaluate the results (Sections 5.6 and 5.7).
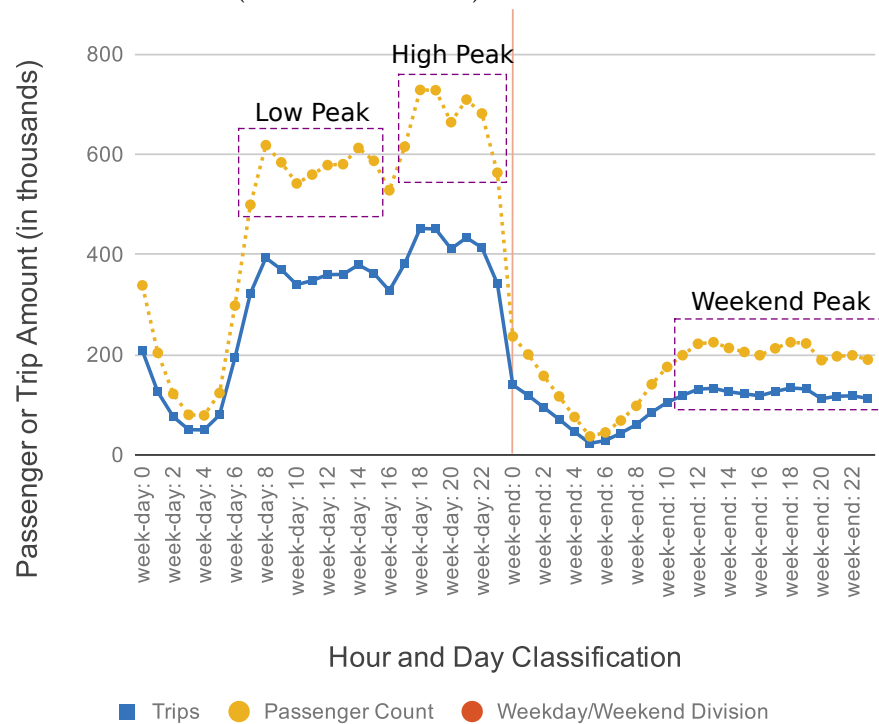

Figure 4. Distribution of trips and passengers in weekdays and weekends per hour; 48 classes were created for every hour of the day in weekdays and weekends [7].

5.3. Data Reduction

We designed a flow clustering technique described in Section 3.3 to reduce the data to be evaluated and keep the relevance of the analysis. Its final step consists of classifying flows between the functional zones. We use a derivative of the curvature equation for this classification. Figure 5 shows the curve of the magnitude of the flows. We smooth the Original Data line, producing the Smoothed line. Later, we use Equation (1) to produce the Curvature line in Figure 5. Its derivative was used to evaluate the maximum, i.e., the Selected Point for the classification (vertical dashed line). Flows with a magnitude above it are considered trending, while the remaining are secondary.
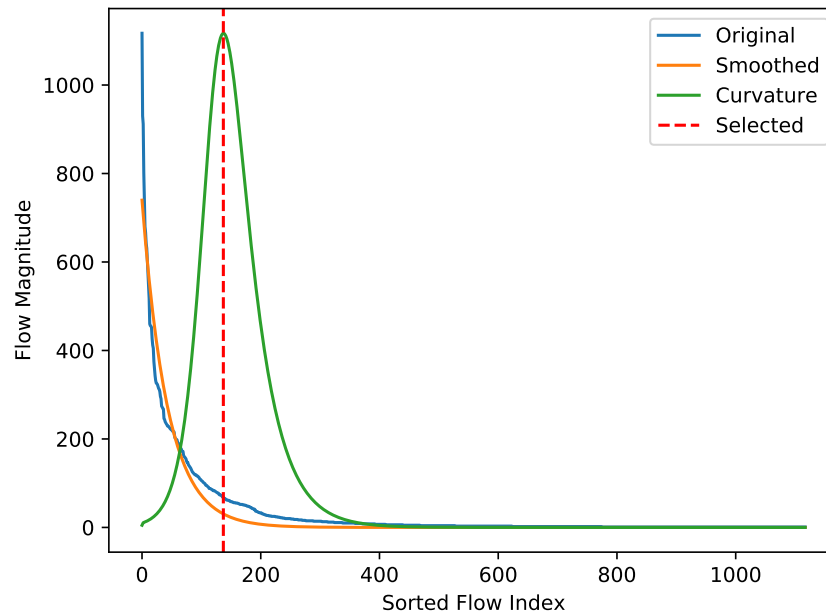
Figure 5. Selection of classification threshold using the derivative of the curvature equation.

Figure 6 shows the classified clustered flows for datasets with and without geo-coordinates. Figure 6a [7] uses the first variation of the flow clustering technique, leading to clusters with a deformed-looking shape. This shape is due to their formation made out of coordinates that were collected from the performed trips. These clusters were obtained from the Yellow Taxis dataset 2016 when geo-coordinates for trip pickup and drop-off points were still available. The clusters shown in Figure 6b,c were tailored—in the anonymization process before data releasing—according to the shapes of some districts and neighborhoods of the city, thus leading to more regular shapes.

Figure 6b,c show a difference in the locations of the departures and arrivals clusters. Yellow Taxis have a historical trend to attend users in Manhattan, while Green Taxis were created to serve the more peripheral zones, being prohibited from starting trips in Manhattan. The focus on covering Manhattan is not strongly present in this sample of flows shown in Figure 6a. Green Taxis are also prohibited from taking passengers to airports, while Yellow Taxis are not, which results in a flow from the bottom-right corner of Figure 6a (where JFK Airport is located) to a cluster near the Brooklyn neighborhood. This is another trend of the Yellow Taxis.

The geo-coordinates allow us to obtain more relevant flows. In contrast, the flows that were obtained with the anonymized dataset resulted in many flows starting and finishing in the same regions. The flows that were obtained in this data reduction phase were used to evaluate the impacts of the hybrid routing, which are described in Section 5.6.
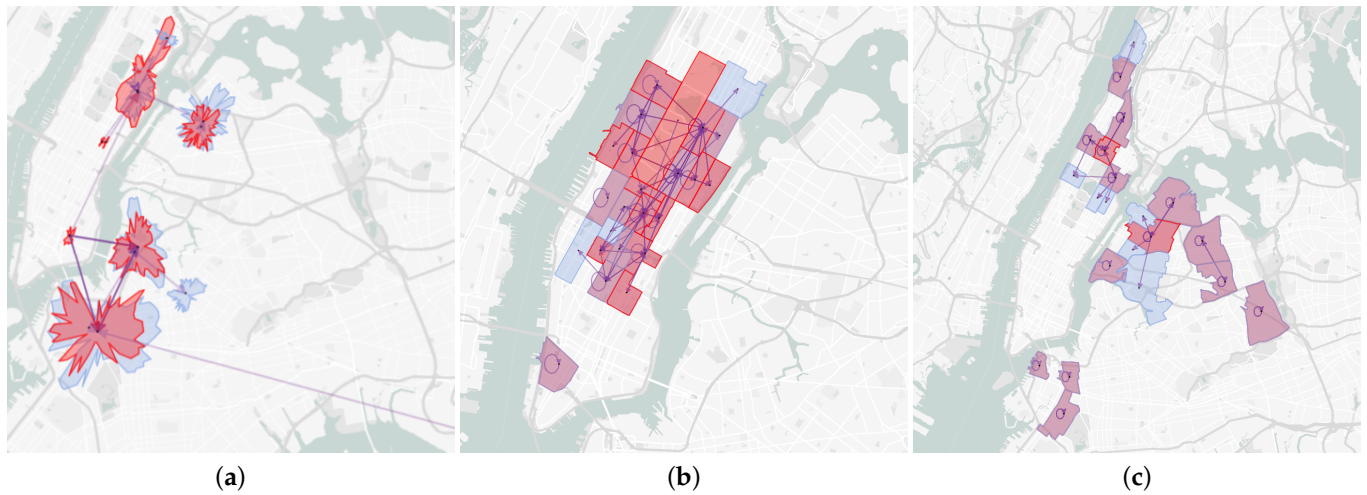
Figure 6. Clustered flows from Yellow and Green Taxis datasets. (**a**) Yellow Taxis with geo-coordinates. (**b**) Yellow Taxis without geo-coordinates. (**c**) Green Taxis without geo-coordinates.

5.4. Clustering Evaluation

In this section, we compare our proposed clustering methodology with the SSTFC [14]. The objective of applying the clustering in the present study is to allow the identification of the most relevant mobility flows, i.e., groups of trips with similar origin-destination that happen with a high frequency. Figure 7 shows the distribution of clusters given the amount of trips that they contain (their size) excluding trips considered as noise. The clusters were divided into classes with sizes greater than 1000, between 100 and 1000, between 10 and 100, and smaller than 10. We show results for two configurations of each approach, using $\alpha \in \{0.5, 0.05\}$ for SSTFC, and using Polynomial and Exponential model to classify the most relevant flows in our proposal, OD-MFC. The evaluation was performed with all of the clusters from the 48 $\langle weekday, hour \rangle$ classes in the non-anonymized NYC Green Taxis dataset.
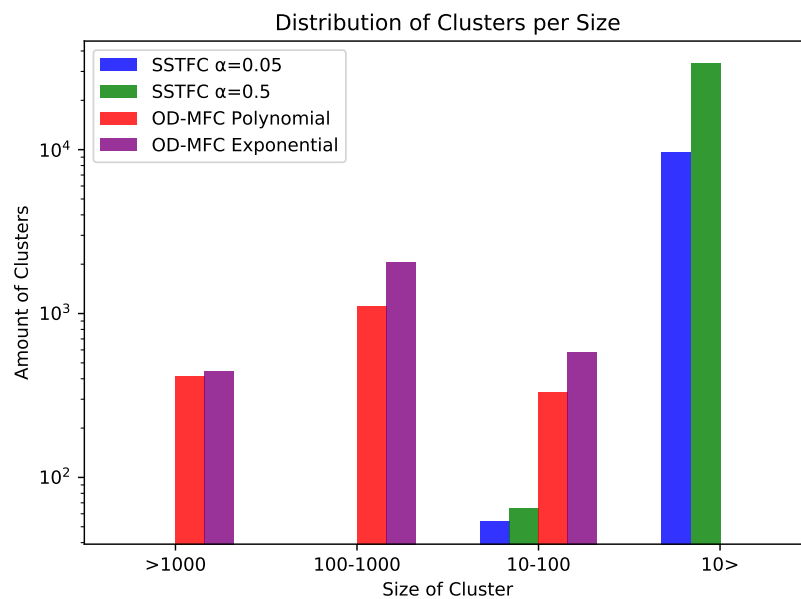

Figure 7. Distribution of flow clusters according to their sizes.

The SSTFC algorithm showed a trend towards identifying a big amount of small clusters, with less than 1% of clusters sized greater than 10 for both parameters studied.

For both scenarios, SSTFC classified more than 99% of the data as noise. OD-MFC could identify more relevant clusters, with all of the selected clusters containing more than 10 trips, and around 60% of the clusters sized between 100 and 1000. OD-MFC classified 71% and 68% of the data as noise, respectively, for Polynomial and Exponential models. In total, we observed 9.7 and 33.4 thousand clusters for SSTFC with α equals to 0.05 and 0.5; and, 1.8 and 3.0 thousand clusters for OD-MFC using Polynomial and Exponential models. Given our objective of selecting relevant flows for evaluation, OD-MFC performed better. Not only more relevant clusters were identified, but also a smaller part of the data was filtered out as noise. There is the possibility that better parameter selection could lead SSTFC towards better results, yet increasing the value of α from 0.05 to 0.5 only increased the amount of smaller clusters identified. Furthermore, initial experiments with smaller α (e.g., 0.005 and 0.0005) did not result in cluster formation, reducing α increases the amount of data classified as noise, as expected. After the clustering step, the average flow of each cluster identified, using the exponential model, was evaluated according to our framework, as described in the following sections.

5.5. Active Transportation Acceptability

The bikeability was obtained using Citi Bike NYC data, as discussed in Section 4.3. We collected data from trips during 2017 and created a cumulative normalized distribution function of the trip lengths (see Figure 8). The curve depicts the probability of some users to use a bicycle given the length of a trip. As its length rises, the number of users drops. This curvature is used as the depreciation factor, in terms of length of trips, when assessing the satisfaction of a given trip segment using bicycles. It is reasonable to use this data since we are using this service as a bicycle provider when evaluating multimodal routes using bike-sharing, i.e., we consider their bike dockers as pickup/drop-off locations when routing trips. Additionally, Figure 8 shows the walkability curve that was used in our work.
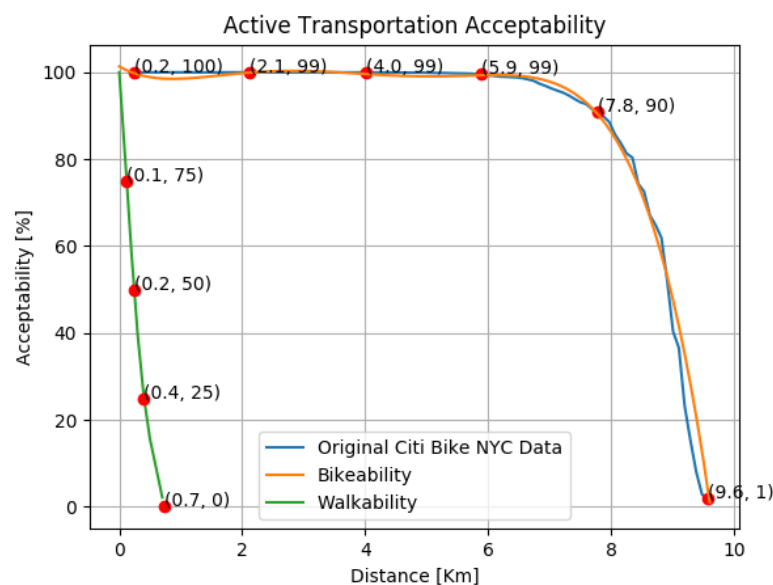

Figure 8. Normalized cumulative distribution function for bike trips using the Citi Bike NYC service in 2017.

To obtain the Original Data curve in Figure 8, trips starting and finishing at the same docker station are dropped, since the bicycles are not being used for mobility. To interpolate the curve and produce Equation (15), we used the highlighted data points (see the bikeability curve in Figure 8). The curve was interpolated as a five-degree polynomial function, whose outcome is a depreciation factor (interval $[0, 1]$) that multiplies the perceived time when using this travel mode. In the figure, the values start to drop fairly around 6 km; before, the impact of the bikeability in the trip's enjoyment is minimal. As shown, the

walkability (i.e., Equation (14)) drops quicker than bikeability, as expected due to the use of bicycles to travel longer distances.

5.6. Hybrid Multimodal Routing

The models described in Section 4 were used to compare the route options for the trip flows that were identified according to the data reduction process. Filtering the trip flows to contain only the analyzed traffic scenarios (i.e., Low Peak, High Peak, and Weekend Peak), a total of 1654 trip flows were evaluated. Table 2 shows an overview of the characteristics of these scenarios, with the averages of trips/flow, duration, and length. The results were obtained for the three main outcomes of the model described in Section 4: (i) Generalized Time, (ii) Generalized Cost, and (iii) Opportunity Cost of Generalized Cost. Figure 9a,b show the modes Uber and Transit mean trips that were performed using those modes, while the Hybrid trips consider the mixture of these modes and the use of bicycles.

Table 2. Average flow characteristics for each scenario.

	Trips/Flow	Duration	Length
High Peak Weekdays (18:00–24:00)	5780	13.6 min	1.87 km
Low Peak Weekdays (08:00–17:00)	6453	13.7 min	2.03 km
Weekend Peak Weekends (11:00–02:00)	2378	15.6 min	2.32 km

Figure 9a shows the distribution of the Generalized Time for the main routing options in the three evaluated traffic scenarios. The Generalized Time is a User Experience metric, which outputs an amount of time that the users would perceive passed on average. In order to compute this metric, we use level-of-service and acceptability multipliers to the actual time elapsed. This way, we aim to convert all times elapsed in different modes to a single base in-vehicle time (IVT). Figure 9a shows that the QoE for hybrid routes obtained the best (i.e., smaller) values of medians. This means that, for most trips, the best option is the Hybrid; also, we observe that the distribution of the values is more concentrated for the Hybrid options, meaning more stability on the results. We also note that the variation of the obtained values increased according to the scenario; this is probably due to the resulting average trip length being greater for Low Peak and Weekend Peak than the High Peak.
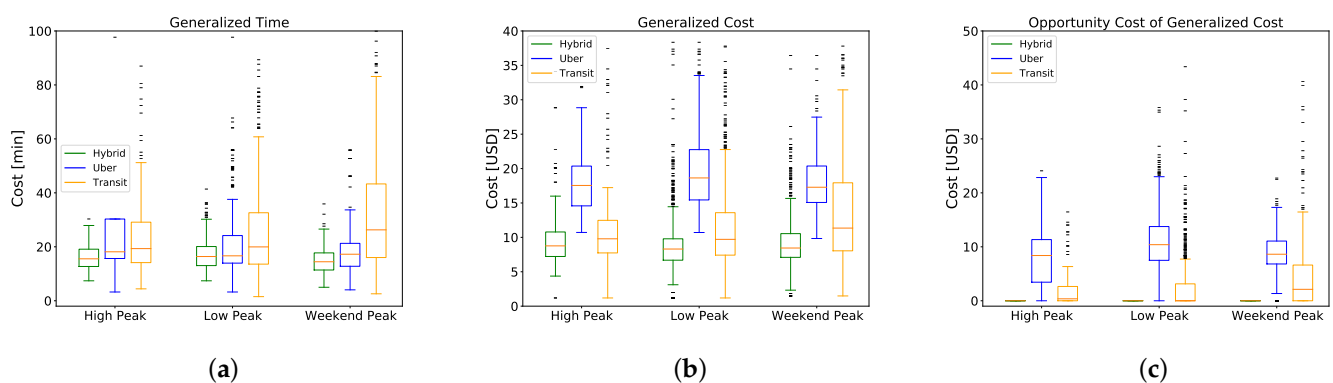

Figure 9. Relative impact of different users' profiles in the route evaluation. (**a**) Distribution of the Generalized Time of trips. (**b**) Distribution of Generalized Cost of trips. (**c**) Distribution of the Opportunity Cost.

The Generalized Time is used to evaluate the Generalized Cost of a trip (see Figure 9b). The generalized time is used with the average income per hour of full-time year-round Americans [45]. This generalized time is added to the price of the trip to obtain its Generalized Cost (includes price, duration, and user experience). The Hybrid option is the best in terms of Generalized Cost. Because Uber options are more expensive, their values are higher than the Public Transportation ones. We observe the values of Hybrid to be under

Public Transportation. This behavior is probably due to the number of bicycle trips that are cheaper than Public Transportation in New York. When considering the Generalized Cost medians, the average saving values are around USD 10.00, when comparing Hybrid with Uber, and USD 2.50 as compared to Public Transportation in all traffic scenarios.

Figure 9c shows the distribution of Opportunity Cost of choosing an option of a trip w.r.t. the Generalized Cost. This data measures the amount of loss for the decision-maker due to selecting a trip among the available ones. A traditional option is an instance of Hybrid routing. Thus, invariably is the best option in this set, leading the Opportunity Cost to always be zero. The traditional options are not necessarily the best routes, which lead to the distribution of Uber and Public Public Transportation in Figure 9c. Additionally, we consider that bicycle options are in the Hybrid group, causing a rise in the values for the common modes. Regarding the transportation mode changes, the best options of the proposed approach have two changes at maximum—with averages number of changes in each ⟨*weekday*, *hour*⟩ class under 0.2.

5.7. Route Selection Based on Profiles

In this work, we propose using the Profile Operator (see Section 4.5) to apply our model in individual route recommendations. This operator combines the attention that us given by a user to the experience or price-related metrics. Figure 10 shows the influence of the profiles for the three already described traffic scenarios. In a real-world system, the profile operator should be calibrated with usage data by applying different learning techniques, such as the one that was proposed in the FAVOUR algorithm [5]. Great variations to the duration metric can be observed in Figure 10, this happens since trips with different lengths, which take varied time to be performed, are evaluated. Conversely, the price metric does not vary much, because, for most of the transport modes, the duration does not have a direct impact on the price (e.g., bus, walk, and the initial hiring fee for Ubers).
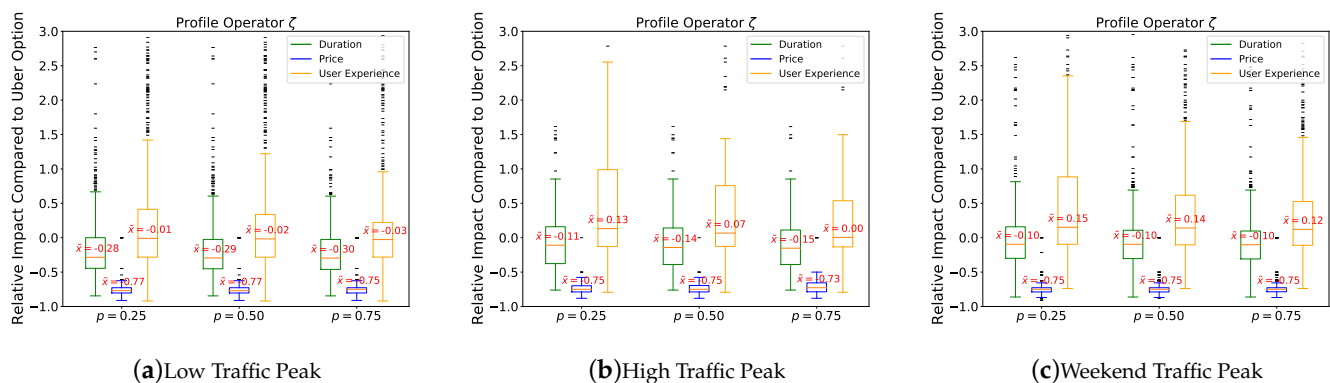

Figure 10. Relative impact of different users' profiles in the route selection.

The profile operator fuses the utility functions after the normalization using a z-score. Next, the ζ profile operator is evaluated, i.e., the route option with the smallest value for each trip is selected for the evaluation—omitting the Uber option, which is used as the baseline. We computed the difference for every metric and then normalized it (divided by the value of the Uber option). We use Equation (19) to calculate the impact in trips, where I is the impact, r' is the Uber option for the trip, and $f(r)$ is the cost function studied (i.e., c_r, t_r, or $T(r)$).

$$I = \frac{\min\limits_{\forall r \in R - \{r'\}} f(r)}{t_{r'}}.$$ (19)

For all of the traffic scenarios, we present the expected results. When p increases, the duration and user experience drop, and the cost values rise. This reflects the idea that users

expecting a better QoE might pay more for the trip and vice versa. Yet, there are specific situations where the price and duration of trips are reduced together, which results in a cheaper trip with better QoE. For instance, when there is congestion in the streets, leaving a vehicle and choosing a different mode, like a bicycle, may result in a faster trip; since the usage of a hired vehicle is reduced, and the price of the trip decreases.

The median values were plotted over the graphs to facilitate the visualization of this behavior. Most of the observations for all scenarios lie between −1 and 1 (81.7% of the evaluated flows), which means that there are no high gains or losses. All of the median is under the zero mark for the low peak scenario, meaning that gains were obtained in more than 50% of the trips for all metrics. In the High and Weekend peak scenarios, only the metrics duration and price had the same performance, while the median for user experience was a little above the zero mark, even though these median values were not too high.

6. Final Remarks

In this work, we presented a novel flow clustering technique for geo-located and anonymized data and a user experience model for urban trips. These are the main contributions of this paper. We detailed the model and the experiments, so it is possible to replicate this work with other datasets (using the code on Github). Overall, this study showed that hybrid routing is a valuable option for traditional urban routing systems.

The proposed model has some parameters that might be, at times, difficult to calibrate. Some of them can be obtained from datasets that one might have access to, such as the average income of users of public transportation and the bikeability and walkability functions; on the other hand, others cannot be easily obtained, such as the weight for the profile operator. A possible solution, as well as a possible future direction of this study, is to use learning techniques to adapt the weights in the profile operator [5]. Some of the issues were identified when running the experiments, such as the anonymized datasets favored the formation of small-length trips. Thus, some modes have more advantages than others in the analysis. However, we presented an approach to cluster data without anonymization. Besides, the Hybrid option showed to be a viable choice for short trips, and previous results [7] showed that it presents good efficiency in more extended trips. Our solution does not explore the full road network when evaluating routes, but a reduced network suggested by services, such as TomTom Routing and Google Directions. We need to check whether the usage of the full graph (or semi-full depending on the heuristic) would bring gains for our work. This could be achieved by changing the MLGLS layer to another, which implements algorithms, such as the shortest path.

One interesting possible continuation to the present study would be to combine the route recommendation with a central ITS of a smart city. For instance, the mode transition points in our proposal depend on the location of the bus terminals and bike dockers. Once integrated with the central ITS, our approach could be used as a evaluation function to optimize the placement of these stations. To put these ideas in practice, our study would have to be integrated with optimization algorithms with the same purpose [46]. Furthermore, in a integrated ITS scenario, our approach could make use of different strategies to not only keep away from traffic jams, but also avoid their creation. For this, different traffic control mechanisms should be integrated with our method, such as traffic lights and street crossings orchestration control [47].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at: https://opendata.cityofnewyork.us/, https://developer.tomtom.com/, https://www.citibikenyc.com/, https://www1.nyc.gov/, https://developers.google.com/, and https://developer.uber.com/ all accessed on 14 May 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Soriano, F.R.; Samper-Zapater, J.J.; Martinez-Dura, J.J.; Cirilo-Gimeno, R.V.; Martinez Plume, J. Smart Mobility Trends: Open Data and Other Tools. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 6–16. [CrossRef]
2. Zhang, D.; Huang, J.; Li, Y.; Zhang, F.; Xu, C.; He, T. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, Maui, HI, USA, 7–11 September 2014; pp. 201–212. [CrossRef]
3. Sweeney, S.; Ordóñez-Hurtado, R.; Pilla, F.; Russo, G.; Timoney, D.; Shorten, R. A Context-Aware E-Bike System to Reduce Pollution Inhalation While Cycling. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 704–715. [CrossRef]
4. Quercia, D.; Schifanella, R.; Aiello, L.M. The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. In Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago, Chile, 1–4 September 2014; pp. 116–125. [CrossRef]
5. Campigotto, P.; Rudloff, C.; Leodolter, M.; Bauer, D. Personalized and Situation-Aware Multimodal Route Recommendations: The FAVOUR Algorithm. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 92–102. [CrossRef]
6. Houshmand, A.; Cassandras, C.G. Eco-Routing of Plug-In Hybrid Electric Vehicles in Transportation Networks. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 1508–1513. [CrossRef]
7. Rodrigues, D.O.; Fernandes, J.T.; Curado, M.; Villas, L.A. Hybrid Context-Aware Multimodal Routing. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2250–2255. [CrossRef]
8. Train, K.E. *Discrete Choice Methods with Simulation*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
9. Dotoli, M.; Zgaya, H.; Russo, C.; Hammadi, S. A Multi-Agent Advanced Traveler Information System for Optimal Trip Planning in a Co-Modal Framework. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2397–2412. [CrossRef]
10. Hong, Z.; Chen, Y.; Mahmassani, H.S. Recognizing Network Trip Patterns Using a Spatio-Temporal Vehicle Trajectory Clustering Algorithm. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 2548–2557. [CrossRef]
11. Pang, J.; Huang, J.; Yang, X.; Wang, Z.; Yu, H.; Huang, Q.; Yin, B. Discovering Fine-Grained Spatial Pattern From Taxi Trips: Where Point Process Meets Matrix Decomposition and Factorization. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3208–3219. [CrossRef]
12. Zhong, G.; Wan, X.; Zhang, J.; Yin, T.; Ran, B. Characterizing Passenger Flow for a Transportation Hub Based on Mobile Phone Data. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1507–1518. [CrossRef]
13. El Mahrsi, M.K.; Côme, E.; Oukhellou, L.; Verleysen, M. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 712–728. [CrossRef]
14. Yao, X.; Zhu, D.; Gao, Y.; Wu, L.; Zhang, P.; Liu, Y. A Stepwise Spatio-temporal Flow Clustering Method for Discovering Mobility Trends. *IEEE Access* **2018**, *6*, 44666–44675. [CrossRef]
15. Birant, D.; Kut, A. ST-DBSCAN: An Algorithm for Clustering Spatial–temporal Data. *Data Knowl. Eng.* **2007**, *60*, 208–221. [CrossRef]
16. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996; pp. 226–231.
17. Amar, H.M.; Drawil, N.M.; Basir, O.A. Traveler Centric Trip Planning: A Behavioral-Driven System. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1521–1537. [CrossRef]
18. Ben-Akiva, M.; Bierlaire, M. Discrete Choice Models with Applications to Departure Time and Route Choice. In *Handbook of Transportation Science*; Hall, R.W., Ed.; Springer: New York, NY, USA, 2003; pp. 7–37. [CrossRef]
19. Hrnčíř, J.; Žilecký, P.; Song, Q.; Jakob, M. Practical Multicriteria Urban Bicycle Routing. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 493–504. [CrossRef]
20. Campello, R.J.G.B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data* **2015**, *10*, 1–51. [CrossRef]
21. Thorpe, J.A. Curvature of Plane Curves. In *Elementary Topics in Differential Geometry*; Springer: New York, NY, USA, 1979; pp. 62–67. [CrossRef]
22. McInnes, L.; Healy, J. Accelerated Hierarchical Density Based Clustering. In Proceedings of the IEEE International Conference on Data Mining Workshops, New Orleans, LA, USA, 18–21 November 2017; pp. 33–42. [CrossRef]

23. Ueda, K.; Yamashita, N. On a Global Complexity Bound of the Levenberg-Marquardt Method. *J. Optim. Theory Appl.* **2010**, *147*, 443–453. [CrossRef]

24. Lorimer, T.; Held, J.; Stoop, R. Clustering: How Much Bias Do We Need? *Philos. Trans. R. Soc. A* **2017**, *375*, 1–18. [CrossRef] [PubMed]

25. Latapy, M.; Fiore, M.; Ziviani, A. Link Streams: Methods and Applications. *Comput. Netw.* **2019**, *150*, 263–265. [CrossRef]

26. Bentley, J.L. Multidimensional Binary Search Trees Used for Associative Searching. *ACM Commun.* **1975**, *18*, 509–517. [CrossRef]

27. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer Networks. *J. Complex Netw.* **2014**, *2*, 203–271. [CrossRef]

28. Sharifzadeh, M.; Shahabi, C. VoR-Tree: R-trees with Voronoi Diagrams for Efficient Processing of Spatial Nearest Neighbor Queries. *Proc. VLDB Endow.* **2010**, *3*, 1231–1242. [CrossRef]

29. Zheng, Y.U. Trajectory Data Mining: An Overview. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 1–41. [CrossRef]

30. Rodrigues, D.O.; Boukerche, A.; Silva, T.H.; Loureiro, A.A.F.; Villas, L.A. SMAFramework: Urban Data Integration Framework for Mobility Analysis in Smart Cities. In Proceedings of the 20th ACM Int'l Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems, Miami, FL, USA, 21–25 November 2017; pp. 227–236. [CrossRef]

31. Buchanan, J.M. Opportunity Cost. In *The World of Economics*; Eatwell, J., Milgate, M., Newman, P., Eds.; Palgrave Macmillan: London, UK, 1991; pp. 520–525. [CrossRef]

32. Bruzelius, N.A. Microeconomic Theory and Generalised Cost. *Transportation* **1981**, *10*, 233–245. [CrossRef]

33. Wardman, M.; Toner, J. Is Generalised Cost Justified in Travel Demand Analysis? *Transportation* **2020**, *47*, 75–108. [CrossRef]

34. Cherlow, J.R. Measuring Values of Travel Time Savings. *J. Consum. Res.* **1981**, *7*, 360–371. [CrossRef]

35. Abrantes, P.A.; Wardman, M.R. Meta-analysis of UK Values of Travel Time: An Update. *Transp. Res. Part A Policy Pract.* **2011**, *45*, 1–17. [CrossRef]

36. Ye, R.; Titheridge, H. Satisfaction with the Commute: The Role of Travel Mode Choice, Built Environment and Attitudes. *Transp. Res. Part D* **2017**, *52*, 535–547. [CrossRef]

37. Moura, F.; Cambra, P.; Gonçalves, A.B. Measuring Walkability for Distinct Pedestrian Groups with a Participatory Assessment Method. *Landsc. Urban Plan.* **2017**, *157*, 282–296. [CrossRef]

38. Nielsen, T.A.S.; Skov-Petersen, H. Bikeability—Urban Structures Supporting Cycling. Effects of Local, Urban and Regional Scale Urban form Factors on Cycling from Home and Workplace Locations in Denmark. *J. Transp. Geogr.* **2018**, *69*, 36–44. [CrossRef]

39. Schaap, N.; Harms, L.; Kansen, M.; Wüst, H. *Cycling and Walking: The Grease in Our Mobility Chain*; KiM Netherlands Institute for Transport Policy Analysis: Den Haag, The Netherlands, 2016.

40. Yang, Y.; Diez-Roux, A.V. Walking Distance by Trip Purpose and Population Subgroups. *Am. J. Prev. Med.* **2012**, *43*, 11–19. [CrossRef]

41. Ryus, P.; Danaher, A.; Walker, M.; Nichols, F.; Carter, B.; Ellis, E.; Cherrington, L.; Bruzzone, A. Quality of Service, Ridership and Service Cost. In *Transit Capacity and Quality of Service Manual*; Transit Cooperative Research Program: Washington, DC, USA, 2013.

42. De Souza, A.M.; Braun, T.; Villas, L. Efficient Context-Aware Vehicular Traffic Re-Routing Based on Pareto-Optimality: A Safe-Fast Use Case. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2905–2910. [CrossRef]

43. Kraus, S.; Arkin, R.C. *Strategic Negotiation in Multiagent Environments*; MIT Press: Cambridge, MA, USA, 2001.

44. Briggs, R. Normative Theories of Rational Choice: Expected Utility. In *The Stanford Encyclopedia of Philosophy*; Zalta, E., Ed.; Center for the Study of Language and Information (CSLI): Stanford, CA, USA, 2017.

45. Survey, A.C. Selected Economic Characteristics: 2013–2017 ACS 5-Year Estimates, 2018. Available online: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2017/5-year.html (accessed on 12 May 2021).

46. Arabi, M.; Beheshtitabar, E.; Ghadirifaraz, B.; Forjanizadeh, B. Optimum Locations for Intercity Bus Terminals with the AHP Approach—Case Study of the City of Esfahan. *Int. J. Environ. Ecol. Eng.* **2015**, *9*, 545–551.

47. Munoz-Organero, M.; Ruiz-Blaquez, R.; Sánchez-Fernández, L. Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on GPS traces while driving. *Comput. Environ. Urban Syst.* **2018**, *68*, 1–8. [CrossRef]