Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 2145716, 14 pages
https://doi.org/10.1155/2021/2145716

Research Article

The Impact of Fleet Coordination on Taxi Operations

Claudio Ruch,[1] **Sebastian Hörl**,[2,3] **Joel Gächter**,[1] **and Jan Hakenberg**[1]

[1]*Institute for Dynamic Systems and Control, ETH Zürich, Switzerland*
[2]*Institute for Transport Planning and Systems, ETH Zürich, Switzerland*
[3]*Institut de Recherche Technologique SystemX, Palaiseau, France*

Correspondence should be addressed to Sebastian Hörl; sebastian.horl@irt-systemx.fr

Received 30 July 2021; Revised 6 October 2021; Accepted 12 November 2021; Published 24 December 2021

Academic Editor: Elżbieta Macioszek

On-demand mobility has existed for more than 100 years in the form of taxi systems. Comparatively recently, ride-hailing schemes have also grown to a significant mode share. Most types of such one-way mobility-on-demand systems allow drivers taking independent decisions. These systems are not or only partially coordinated. In a different operating mode, all decisions are coordinated by the operator, allowing for the optimization of certain metrics. Such a coordinated operation is also implied if human-driven vehicles are replaced by self-driving cars. This work quantifies the service quality and efficiency improvements resulting from the coordination of taxi fleets. Results based on high-fidelity transportation simulations and data sets of existing taxi systems are presented for the cities of San Francisco, Chicago, and Zurich. They show that fleet coordination can strongly improve the efficiency and service level of existing systems. Depending on the operator and the city's preferences, empty vehicle distance driven and fleet sizes could be substantially reduced, or the wait times could be reduced while maintaining the current fleet sizes. The study provides clear evidence that full fleet coordination should be implemented in existing mobility-on-demand systems, even before the availability of self-driving cars.

1. Introduction

Once installed and set up, robots typically complete tasks in a steady quality and at unmatched speed. Their operating cost is relatively small compared to the amount of labor they perform. They have virtually unlimited operating hours. If a robot once does fail, it can be immediately replaced with an almost identical machine without time-consuming training. These are some of the reasons why automation, that is, replacing and supplementing humans with machines, has profound implications on the profitability and productivity of businesses. Its influence on employment and the structure of society is not less profound [1]. One side effect of automation is that it may change the fundamental setup and organizational paradigms of an industry. We consider transportation and specifically the emerging mode of on-demand or shared mobility. The term "shared mobility" typically bundles several modes of transportation that are neither the privately owned motorized car nor the conventional public transit on fixed lines and schedules. Shared mobility includes for instance one- and two-way bike- and car-sharing schemes [2, 3], ride-hailing schemes [4], or full-fledged mobility-on-demand systems with self-driving cars [5]. Today's most prevalent form of on-demand mobility is taxi schemes. They have been present in cities for decades [6] with only minor changes in the way they are operated. However, at this point, robotic cab drivers in the form of self-driving cars are about to start operation [7] and promise to rapidly induce major changes. The principal reason is that robotic taxicabs have a very different behavior compared to human drivers. Robots will not have individual preferences, for example, a preferred area of the city to roam in and search for waiting customers. They will have almost unlimited operating hours, and their number can be freely adjusted without labor market restrictions. An additional important feature is that the behavior of robotic cars will be fully determined by the operator. Thus, self-driving cars will immediately overturn the paradigm that each taxi decides on its own schedule, that is, where to wait

for new customers, when to enter and leave a work shift, and so on. A taxi system using robotic cars is coordinated *by design*.

A coordinated system can potentially reach a system or social optimum, while a system with selfish and strategic drivers in general cannot [8]. Therefore, the transition of taxi scheme operation to coordinated fleets is expected to improve overall performance. In this work, we attempt to quantify these effects. Our objective is to understand precisely how large the gains in service level and efficiency would be if today's taxi systems were operated in a coordinated manner. Interestingly, even without the availability of fully autonomous vehicles, there are no technical limitations that inhibit the implementation of these concepts in today's taxi systems. Taxi system operators as well as ride-hailing companies such as Lyft and Uber would though have to change their business model and way of remunerating the drivers to harvest the benefits of full fleet coordination.

In order to quantify the impact of fleet coordination, we proceed as follows: first, we take recordings of existing taxi schemes and analyze them to quantify the status quo. Then, we create simulation scenarios in the framework AMoDeus [9] that accurately replicate the conditions of the real world. Then, we serve the same demand under the same conditions with a hypothetical coordinated mobility-on-demand system. For this, we use different state-of-the-art fleet operational policies and varying fleet sizes and show operating points in comparison to the taxi system.

We complete this assessment for the cities of San Francisco, Chicago, and Zurich for which we have taxi data available. The used simulation scenarios and the source code are available online and thus allow rapidly evaluating other cases in future research.

The paper is organized as follows. In Section 2, we first provide a detailed summary of the work related to our paper, covering insights from both the economic and engineering research domains, and we position the contributions of our paper relative to them. After, we present the used methods and simulation models in Section 3. Specific case studies of the taxi systems in the cities of San Francisco, Chicago, and Zurich are presented in Section 4. The city-specific results are summarized and compared in Section 5. Finally, Section 6 provides a general summary of conclusions.

2. Related Research

In this work, we aim to estimate the gains in service level and efficiency resulting from taxi fleet coordination. Different preceding studies have quantified the inefficiencies of existing taxi systems:

2.1. Data Analysis Approaches. One important class of contributions focuses on the in-depth analysis of taxi trace or trip data sets: in [10], a taxi data set recorded in Singapore during a 24-hour period with a fleet of 6,230 taxis of the operator ComfortDelGro is analyzed to find out whether the available taxis are used effectively to cover unmet demand. The authors rightfully argue that neither high wait times alone nor low taxi utilization rates can be interpreted as a sign of inefficiency. In contrast, a simultaneous occurrence of empty taxis (low fleet utilization) and long waiting times (low service level) indicates an inefficient taxi allocation. Therefore, if wait times are unusually high for a given level of occupancy or if occupancy is unusually low for a given level of wait times, the system performance is suffering from an inefficient assignment of demand and supply. The data analysis reveals several such situations of inefficient demand and supply assignment, for example, during the morning rush hours in the central business district or during late-night hours.

A similar study is [11], in which a 2-week taxi trajectory data set of 14,000 taxis in the city of Chengdu, China, is considered. The authors measure that taxis are productive about 59% of the time and conclude that the number of operating taxis in the city is reasonable. Nevertheless, they detect substantial imbalances between different zones of the city and suggest that a better mechanism to balance taxis between the geographic areas of the city should be implemented.

A somewhat different approach is taken by Cramer and Krueger [12]. In this work, the efficiency of the ride-sharing service Uber to that of taxi drivers is compared by assessing their time capacity utilization rate and distance capacity utilization rate in the US cities of Boston, Los Angeles, New York, San Francisco, and Seattle. In all cities except New York, Uber drivers show a time capacity utilization rate that is around 40% higher compared to conventional taxi services; their distance capacity utilization rate is between 41% and 59% higher. One of the possible reasons the authors see for Uber's edge on efficiency is "more efficient driver-passenger matching technology," but the hypothesis is not verified. Other possible reasons listed by the authors include the flexible labor supply model of Uber, the utilized surge pricing method, or simply the larger scale. The authors estimate that Uber drivers could charge 28% less than taxis while earning the same hourly revenues. While the study is a motivation for the work at hand, it leaves open several questions: although Uber uses algorithms to improve service level and efficiency, their operation cannot be considered to be fully coordinated. Drivers are left with many degrees of freedom as well as the responsibility for their own profit and loss. Thus, the magnitude of further improvements through fleet coordination remains unclear. Then, the comparison is made in the same cities but not on the same travel demand. It is not possible to distinguish differences in operating efficiency that are due to different operations or due to different demands, for example, it may even be that Uber succeeds in serving a higher fraction of operationally efficient trips at the expense of the taxi operators. The use of surge pricing increases the probability that the ride-hailing companies serve different demands than taxis.

Approaches based on data analysis focus on the identification of inefficiencies in existing taxi systems, but they do not answer the question to what degree these inefficiencies could be resolved with the improved operation and fleet coordination. As an example, the late-night hour inefficiencies detected by Santani et al. [10] might be systemic

and possibly cannot be solved even with the best coordination available for such light-load demand cases [13].

2.2. Taxi Market Models. Next to a pure data analysis approach, some contributions attempt to deepen the understanding of the market dynamics in play. Most of these contributions focus not only on operation but also on the elasticity of the taxi travel demand. In [14], a simplified model of a taxi market is presented with one single journey origin in order to explain the interplay between taxi utilization, expected wait time or taxi availability, demand, and expected earnings. The model explains convincingly that an increasing demand may simultaneously increase both utilization and availability as more taxi drivers can generate sufficient profits and remain in operation. Built with one single queue, the model does not consider spatial effects. The spatiotemporal distribution of the demand is simplified to a basic model and does not capture effects present in real taxi demand, including large variations in space and time and correlated requests. The study does not consider effects on vehicle miles traveled. In [15], an elaborate model of taxi markets is proposed in which matching frictions as well as entry regulations are considered. So-called matching frictions occur because an idle taxi and a waiting customer may not be at the same location. Thus, the process governing the search behavior of both taxis and customers determines the extent of these frictions. The authors consider both normal street-roaming search behavior and one form of centralized dispatching in which the closest taxi is assigned to an open request. A key strength of the study is that it relates the type of matching to general market dynamics, for example, their influence on the demand for taxis. The drawback is that several modeling assumptions are made that are potentially oversimplifying in a study focusing on operational efficiency: a highly simplified regular street network and Manhattan distances are used, vehicles drive constant speed, passengers may be lost and exit the system unserved, and vehicle rebalancing is not considered. The choice of modeled variables by Frechette et al. [15] (vehicle utilization, vehicle availability, and earnings) does not allow detailed insights into how the operational policies would influence service level and vehicle miles traveled. Other related studies focusing on the economics of the taxi market are, for instance, [16–18]. Many highlight the impact of regulation, for example, predetermined taxi fares or taxi medallion auctions [19] and are thus only loosely related to the work at hand.

2.3. Making Individual Taxis Succeed: Recommender Systems. Both the analysis of taxi traces and economic models provide evidence that inefficiencies are present in most taxi systems. Based on this observation, a consecutive problem is to study how these inefficiencies could be avoided: in this spirit, the authors in [20] attempt to identify the most successful taxi drivers and their strategies. Based on a study of taxi data from Wuhan, China, they find out that highly successful drivers often drive away from downtown areas and their distribution is "highly correlated with traffic conditions." Although very relevant, the work does not answer the question of how many improvements would be possible on a system level and to what degree fleet coordination of the fleet would be required for it.

Apart from identifying successful drivers, they could also be *created*. This is what a class of publications attempts to do with so-called *recommender systems*, which are designed to influence the decision-making of taxi drivers. These systems use different data as input and typically present recommendations to individual taxi cab drivers and/or passengers. The recommendations should reduce matching frictions. The literature on recommender systems is vast. As an example, the work presented by Yuan et al. [21] is based on a Beijing taxi trace data set. Analysis of the data reveals that many taxis cruise extensively and exhibit high vehicle miles traveled without having high earnings, a clear indicator for inefficiencies. The authors propose two recommender systems as a remedy, one for taxi drivers and one for passengers. The one for taxi cab drivers recommends the parking place with the highest probability of finding a customer and the route to this parking place. The recommender system for customers suggests a new location within walking distance that increases the chance of catching an empty taxi. Although the presented recommender systems are evaluated in simulation and in field trials, the presented results do not allow precisely understanding what their effects on vehicle miles traveled were.

In [22], a recommender system is presented with the ability to recommend a sequence of potential pickup points to a driver to minimize empty travel distance. However, the evaluation of the method does not allow understanding the degree of inefficiency of taxi systems that are currently in operation. Another recommender system presented by Ding et al. [23] does also not consider the entire fleet but only individual taxis.

The recommender system in [24] focuses on the taxi movements between trips, called "cruising." The proposed operational policy called *pCruise* attempts to minimize the cruising distance of taxis by computing cruising routes for them with the highest probability of finding a waiting customer. Its effectiveness is proven in a custom-made and simplified transportation simulator with taxi data from Shenzhen, China. An interesting aspect of the publication is that it considers the case when the strategy is only used by some percentage of the city's entire taxi fleet, that is, by one single operator.

2.4. Dynamic Analysis. One way to have more degrees of freedom in the analysis of taxi systems is to use agent-based transportation simulations, for example, MATSim [25]. Simulators like MATSim allow for the efficient representation of road network dynamics and the consideration of dynamic demand with coevolutionary algorithms and mode choice models.

In a set of agent-based simulation studies [26–28], the authors analyze different operational policies in simulation with fleet sizes between 1,000 and 2,000 vehicles serving trips derived from taxi trace data for Berlin and Barcelona. The results show that different operational strategies are most

suitable depending on the system load. The studies show that upscaling of the demand is necessary to observe a substantial impact on waiting times for coordinates fleets.

The study presented by Poulhès and Berrada [29] implicitly highlights the difference between uncoordinated and coordinated mobility-on-demand systems by comparing one instance of a local and one instance of a global operational policy. The results focus on customer service quality, that is, wait time, and show that it is improved through centralized coordination.

Other studies based on multiagent simulation have been presented. However, most studies are performed with autonomous mobility-on-demand (AMoD) systems in mind where, based on citywide modeling efforts, a large number of trips performed by car or public transport users is served by an on-demand system [30]. Many of those studies focus on algorithmic developments in terms of fleet operational policies, for example, [31] or [32]. The developed strategies could be applied to real-world taxi systems.

The work that is most similar to our contribution is [33]. The authors present a graph-based approach to analyze potential efficiency gains in an urban taxi service system. Based on their method, they identify that "inefficiency arises largely due to the lack of globally shared information among taxi drivers and passengers." The underlying idea of their approach is to find a more efficient assignment of taxis to requests that guarantees the pickup time and date of the passenger recorded in the original data set. The approach is applied to taxi data published by the New York City Taxi and Limousine Commission and suggests that the taxi idle time can be reduced by 78–90%, the total empty distances by 60–82%, and the total revenue loss by 66–82%. Furthermore, in most cases, about 67% of the original number of taxis would be sufficient to serve the original demand with the same service level. The study provides an upper bound on the possible improvements on idle time, total empty distance, and total revenue for the case of New York. One limiting assumption that is made in the paper is that Manhattan distances are used to simplify calculations. The cost function is a weighted combination of taxi idle time and empty trip distance. Therefore, the service level is not explicitly taken into account and is also not a tunable parameter.

2.5. Summary. In summary, the existing literature differs in several aspects from the work at hand. Data analysis approaches such as [10, 11] detect inefficiencies in existing taxi systems but do not quantify the exact improvements that could be made using state-of-the-art fleet operational policies. Instead, we attempt to quantify exactly how the system metrics of mobility-on-demand systems would change under the influence of coordinated fleet operation.

Studies focused on the modeling of the economic relations in the taxi market, for example, [15], often work on low-resolution or simplified road networks, with vehicles traveling at constant speed in uncongested networks and governed by basic or no coordinated operational policies. In our work, we instead consider a detailed and high-resolution road network with variable speed limits and congestion effects, and we consider a coordinated fleet operation.

Studies presenting recommender systems, for example, [21–23], typically focus on the actions of one single taxi. Instead, we consider the coordinated operation of the entire fleet. Furthermore, recommender systems often focus on one particular metric, for example, vehicle availability. Instead, we evaluate jointly different metrics for the entire fleet. Existing studies in agent-based simulators such as [26–28] do not provide a comparison between a simulated coordinated taxi system and a real existing taxi system as reference data are not available or reported. Empty vehicle repositioning (rebalancing) and vehicle miles traveled are not considered in these studies. Instead, our results do not only focus on customer service quality (wait time) but contrast this metric with the operators' viewpoint, that is, the vehicle distance traveled of the system. We also compare these metrics to real taxi system data.

In comparison to the graph-based approach presented by Zhan et al. [33], our study allows estimating the trade-off between different system metrics. The key difference of our work is, however, that Zhan et al. [33] assume that all taxi trips are served at exactly the same time as in the data set. Instead, we use *online* operational policies that allow modifying wait times of individual trips, that is, we allow for some flexibility in the operation, provided that the same demand as in the real taxi system is covered. With this approach, we can consider all elements of fleet operation including online reassignments of trips and vehicle rebalancing, and we can analyze the full set of possible system set points.

In summary, the results in previous literature do not allow for a definite conclusion of the performance and efficiency gains in taxi services through fleet coordination. In our study, we quantify these effects for the taxi systems of San Francisco, Chicago, and Zurich. To that end, we

(i) Provide a generic methodology for setting up (coordinated) taxi system simulations given a similar class of taxi request information using an open-source framework and method to estimate network travel times

(ii) Show that those analyses can be performed using two open (San Francisco and Chicago) and one proprietary data source (Zurich)

(iii) And analyze the trade-off between wait times and vehicle miles traveled for different fleet sizes and state-of-the-art fleet operational policies

It should be noted that these analyses are data-driven, that is, no synthetic demand estimates are used, and that we compare with real-world system metrics that are based on the obtained trip data.

3. Methods

Among the publicly available taxi data sets, there is no common standard format. Some contain the traces of the taxis, encoded with data objects of varying scope and

recording frequency. Other data sets only contain information on a trip basis: each taxi trip is logged and additional information is recorded, for example, origin, destination, distance, or payment information. However, all analyzed taxi data sets allow extracting the trips on a certain day including their origin, destination, submission time, and duration. Furthermore, they all allow estimating the number of taxis that were used to serve the trips and to determine a lower bound of the empty vehicle distance driven by the fleet. These data points constitute the raw material to create transportation simulation scenarios. The methods used in this process are presented in this section; additional particularities to the data sets are treated separately in Sections 4.1–4.3.

3.1. Comparison of Simulation and Recorded Data. Our objective is to assess whether the operation of existing taxi systems improves with coordinated fleets. One natural way of answering this question would be to find two days of operation with similar characteristics and *actually* operate the fleet in a coordinated manner on one of the days.

Several problems and shortcomings complicate the execution of this idea: the necessary effort to manage the logistics of such an experiment is very large and far beyond the possibilities of most research institutions. Even if these hurdles could be overcome, each pair of days would only allow assessing one system set point.

However, as we explain in Section 3.4, it is necessary to compare many different set points, which are mainly characterized by the chosen fleet size and the operational policy. Then, such experiments would never allow comparing the exact same travel demand that necessarily varies for different days.

For these reasons, we have chosen a simulation approach, which is based on three principles: first, we do not simulate the existing taxi system but only compare information, which can be directly computed from the given data, for example, the vehicle distance traveled. An accurate representation of an existing taxi system in simulation is difficult, as the influence of the human factor is large, for example, every driver is independent to some degree and his personal preferences, and strategies would have to be understood and modeled accurately.

The second guiding principle of our approach is that we replicate the simulated coordinated taxi system as accurately and as replicable as possible: our transportation scenarios and the source code are available online. Finally, we compare conservatively, that is, we always favor methods that evaluate the real taxi system positively. Therefore, our results show a lower bound of the efficiency and service level improvements possible with fleet coordination.

These guiding principles lead to a number of design decisions: we use the queuing-based traffic simulator of MATSim [25], which is a publicly available open-source software that has been tested and validated extensively. Its setup offers both accuracy and fast simulation speeds, which allows conducting parametric studies. The fleet operational policies that we use are also publicly available; they are part of the open-source software package AMoDeus [9] that we have presented in previous work.

We read out the travel demand served by the taxis directly from the taxi data and serve the same trips in the simulation environment, with the identical origin, destination, and submission time.

In order to obtain comparable and consistent results, the modeling of the road network and its traffic condition is important. Details about how this process was carried out are provided in the following sections: the generation of the high-resolution road network is explained in Section 3.2, and the estimation and adjustment of network speeds is explained in Section 3.3.

3.2. Road Network Generation. MATSim road networks are directed graphs composed of nodes and links, which represent queues. Vehicles are always located on links and pass from one link to another at intersections. In order to generate high-resolution road networks efficiently, the data available from OpenStreetMap [34] were used. A MATSim readable network was generated from this data using the generator pt2matsim [35]. For every road segment, the free-flow speeds were set to the speed limits, which is typical for this road segment.

3.3. Estimation of Network Speeds. The network generation captures different types of roads with different free-flow speeds, for example, 30 or 50 miles per hour. However, the taxi data sets were recorded under real operating conditions with congestion and reduced speeds. These network conditions present at the time of the data recording must be replicated in the simulation environment in order to make meaningful comparisons.

In many studies, the problem of network calibration is viewed from a different angle. MATSim studies focused on dynamic demand take as input a population with specific activities, travel plans, and a personal utility function that values time, travel time, punctuality, and so on. Then, based on the coevolutionary algorithm of MATSim, the Wardrop equilibrium [36] of the scenario is estimated, and from it, the traffic speeds can be derived.

In this work, we know the travel times and need to adjust the network speeds to them, which is a different problem. In the typical case, the number of variables is much higher than the number of measurements. For example, in the Chicago scenario of this study, there is a final number of $M = 36\,016$ taxi trips with known duration, but there are approximately 142,118 network links for which the traffic speed must be estimated during each time of the day. Using 7.5 minutes resolution time steps, this results in a total of 27,286,656 degrees of freedom to satisfy these 36,016 constraints. These have to be chosen in a way such that the simulated duration of trips is identical to the one in the original data set.

To estimate the network speeds for this study, a method was implemented that produces an estimate of the traffic speeds on each link and at each time step based on the known trip durations. It takes network links $i \in \{1, \ldots, N\}$ with free-flow (=maximum) speeds f_i at each link. These are

typically speed limits assigned according to the type of road. Furthermore, a set of taxi trips $\{T_1, T_2, \ldots, T_M\}$ with known start and end location in the network, start time, and trip duration $d_{c,j}$ is needed. The trip durations must be longer than free-flow trip durations in the network; otherwise, the trip is not considered valid. Tunable parameters for the algorithm are two real constants ϵ_1 and a time constant τ, a maximum number of iterations $\mathcal{I}_{\max}$, a tolerance ρ, and a cost function $\mathcal{F}$: $\mathbb{R}^M \longrightarrow \mathbb{R}^+$. The parameter ε_1 determines how aggressively the network speeds should be reduced to match the observed trip durations. The parameter ε_2 determines the rate at which random samples are selected; alternatively, the trip with the worst cost is selected. The time constant τ determines the time resolution of the estimated network speeds. $\mathcal{I}_{\max}$ and ρ determine when the algorithm is terminated. $\mathcal{F}$ influences the characteristics of the error distribution. The principle of the algorithm is to iteratively make the travel time slightly longer for trips until the measured trip duration $d_{p,i}$ of the trip in the network matches the recorded trip duration $d_{c,i}$ from the dataset. The absolute difference of trip durations for trip j is denoted as $\lambda_j = |d_{p,j}/d_{c,j} - 1|$. The outputs of the algorithm are estimated network speeds $\{s_{i,t}\}$, for $i = 1, \ldots, N$ and all time steps t. The principle is illustrated in Figure 1.

The algorithm is described in detail in Algorithm 1.

For the scenarios generated in this work, good results were obtained with a small value of $\varepsilon_1 = 0.05$, with $\varepsilon_2 = 0.8$, $\tau = 450s$, $\mathcal{I}_{\max} = 100000$, and $\mathcal{F}$ as the max function.

3.4. Simulation Approach. Existing data sets of taxis are not in a standard format but very different depending on the source. Thus, many of the modeling decisions are explained in Sections 4.1, 4.2, and 4.3. This section summarizes the general approach chosen, which is identical for all cases.

In general, a mobility-on-demand system should use as few vehicles as possible to serve as many requests as possible with minimal wait and journey times and while simultaneously reducing the empty vehicle distance to a minimum. These are conflicting objectives, for example, a reduction of the fleet size will generally result in longer wait times.

Thus, there are several setpoints that an operator might choose, depending on the requirements. Typically, a system setpoint is reached by an appropriate choice of the operational policy guiding the behavior of the vehicles and of the fleet size. Our approach is therefore to choose different operational policies and fleet sizes and visualize the set of all possible setpoints of the coordinated system as a curve. Specifically, for a range of fleet sizes, we record the mean wait time as a metric for the service level and the empty vehicle distance as a metric for operational efficiency. Furthermore, we run all simulations with one of three simulation policies to favor either service quality or operational efficiency. The assessed policies are as follows:

(1) The global bipartite matching policy (GBM) described by Ruch et al. [9] repeatedly solves an Euclidean bipartite matching problem between the location of unserved customer requests and the locations of idle vehicles. This policy effectively minimizes empty vehicle distance.

(2) The fluidic feedforward rebalancing policy (FFR) uses the same matching algorithm as GBM but additionally executes vehicle rebalancing according to the fluidic feedforward method presented by Pavone et al. [37].

(3) The last policy (MFAR) rebalances vehicles according to the model-free adaptive repositioning policy presented by Ruch et al. [38]. This results in significantly reduced wait times but does increase the empty vehicle distance.

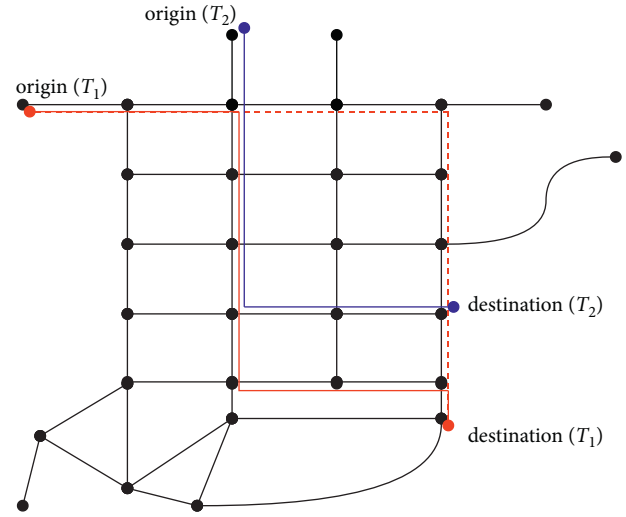

Figure 1: Illustration of Algorithm 1. The red and blue paths are the shortest-duration paths in iteration $\mathcal{I} = 1$; in the next iteration, the red path changes to the dashed line in response to the reduced traffic speeds along the old path.

While the first two algorithms have been included in a parametric simulation study assessing the performance of fleet operational policies by Hörl et al. [32], they have been additionally assessed in comparison to the third strategy in [38]. The approaches have been selected as they have been shown in those studies to differ in operational characteristics. While GBM minimizes the system distance at the cost of wait time, MFAR minimizes wait times at the cost of an increased empty distance. FFR, finally, has been shown to provide a well-balanced trade-off.

The global bipartite matching algorithm is based on Euclidean distances, while the travel time estimated for the MFAR algorithm has been precomputed.

Note that we do not consider algorithms where multiple passengers can be picked up by one vehicle. A large body of literature on respective algorithms exists [39, 40], and numerous studies have been performed where those algorithms are tested in simulation, for example, by [41], who test an algorithm on taxi demand data for Manhattan, or by Bischoff et al. [42], who extend their work on Berlin using a heuristic ride-pooling approach. However, with the possibility to perform ride-pooling, a number of underlying assumptions on the willingness to share are made. As in the

input Tolerance ρ, maximum iterations $\mathscr{I}_{\max}$, $\varepsilon_1 \in [0,1]$, $\varepsilon_2 \in [0,1]$, sampling time τ, set of taxi trips $\{T_j\}$, $j \in \{1,\ldots,M\}$
initialize $\{s_{i,t}\} = f_i \,\forall$ links i, discrete times t, iteration $\mathscr{I} = 0$
initialize a data structure $\mathscr{D}_1$ that sorts trips T_i according to how often they were randomly selected and records λ_i at selection time for each trip
initialize a data structure $\mathscr{D}_2$ that sorts trips T_i according to λ_i
while $\mathscr{I} < \mathscr{I}_{\max}$ and $\omega < \rho$ **do**
with probability ε_1**:** select a trip j randomly from the set of least randomly selected trips in $\mathscr{D}_1$, otherwise select the trip j with highest λ_j from $\mathscr{D}_2$
compute path time $d_{p,j}$ using $\{s_{i,t}\}$, origin and destination of T_j; if a random trip was selected update $\mathscr{D}_1$
compute $r = d_{p,j}/d_{c,j}$ with $d_{c,j}$ recorded duration of T_j
compute $f = \min(1 - (1-r)\varepsilon, 1)$
scale all $s_{i,t}$ in path with f for $t \in [t_{j,\text{start}}, t_{j,\text{end}}]$
compute path time $d_{p,j}$ using updated $\{s_{i,t}\}$, origin and destination of T_j to update $\mathscr{D}_2$
compute the cost ω using $\mathscr{F}$ and the last M randomly selected trips
$\mathscr{I} = \mathscr{I} + 1$
end while
output $\{s_{i,t}\}$

Algorithm 1: Taxi Trip Network Calibration.

present work, we compare empirical taxi data with the wait time as the main indicator for customer experience; ride-pooling is not considered for the sake of consistency.

An operator of a coordinated fleet of taxis would place the taxis strategically at the beginning of the day before morning travel requests begin to enter the system. There is considerable potential to optimize a system with this placing choice, for example, just the expected necessary number of vehicles can be placed in every zone. In this work, we limit ourselves to the most simple of the smart placements for the sake of comparability: all vehicles are randomly distributed on the set of roads on which travel requests arrive during some time of the day.

4. Experiments

In the following sections, simulation experiments for the cities of San Francisco, Chicago, and Zurich are presented. While, here, key performance indicators in terms of empty distance and waiting time are reported for each city and combinations of fleet size and control policy, the results are compared and put into relation in Section 5.

4.1. Taxis in San Francisco. Probably, the most studied data set of taxi traces [43] was recorded in the city of San Francisco and is publicly available at [44]. It contains the traces of a total of 536 taxis that were recorded in the time interval between May 17, 2008, 03:00:04 and June 10, 2008, 02:25:34.

Each trace is composed of measurements that contain a time stamp in Unix Epoch Format, the location latitude and longitude in WGS84 coordinates, and the taxi status $\in\in\{$available, occupied$\}$. The sampling time is irregular and mostly around 1 min. However, sometimes it is much larger. The data set contains a total of 464,045 customer trips.

For the work at hand, we have selected the day June 4. On that day, 18,813 trips were recorded, and 495 different taxis were in operation. Out of this total number of trips, 53 had a delay longer than one hour; 258 were slower than 5 mph; 204 were shorter than 200 m; and 375 did not have a nontrivial path, that is, they started and ended on the same link. These trips were considered invalid recordings and thus removed from the analysis. Combined, 550 trips were removed by the above filters resulting in a final number of 18,263 trips for the day.

In the set of these trips, a total of 5,871 trips is faster than the road network allows for, if the start and end time of a trip is taken directly from the last recorded empty and the last recorded occupied time step. This is caused by large time steps of several minutes up to several hours between some of the measurements and effectively makes it impossible for these trips to know the exact duration. As Algorithm 1 requires a trip duration slower than maximum speeds in the network allow for as a start condition, these trips were not considered for network speed estimation. Instead, the remaining 12,544 trips that are slower than the network are assumed to accurately represent the traffic conditions in the road network on the day the recordings were made.

These trips are used to compute a network speed profile using Algorithm 1. Figure 2 shows the cumulative density function of both the simulated and the recorded trip durations.

The fit is sufficiently good, although the model of the road conditions makes simulated trips slightly slower than recorded ones. As this corresponds well to our principle of comparing conservatively, it is acceptable. The empty vehicle distance of the real taxi system is computed by analyzing all the valid trips and by computing the shortest-time path between trip end locations and subsequent trip start locations.

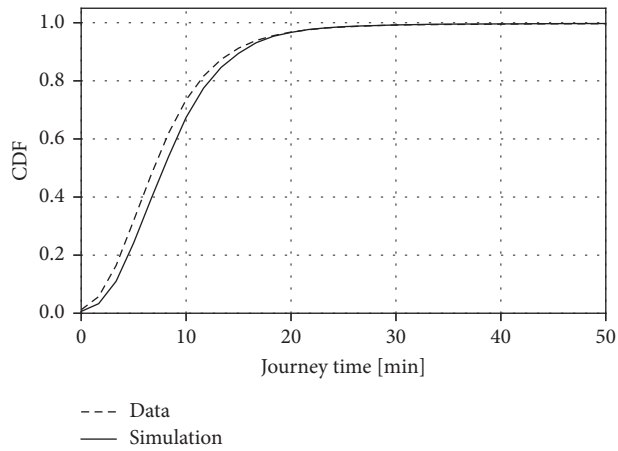

Figure 2: Comparison of journey times in simulation and in the original data set for San Francisco taxi trips.

The sum of these distances can be seen as a lower bound on the empty vehicle distance. During the actual day, vehicles could have taken a less direct path between the locations, spent some idle time cruising in neighborhoods searching for customers, and so on. Unfortunately, the data set does not allow determining the wait time that the taxi customers experienced. An empty distance of 62,907 km is estimated. Having created a suitable representation of the taxi demand and network conditions, we can additionally serve the same demand under the same conditions with different operational strategies and fleet sizes.

The resulting operational characteristics and the range of operating points of the real taxi system are shown in Figure 3. The figure shows the results from a series of simulations where the given set of requests is served by fleets from 400 to 700 vehicles and with the three operational policies. Each simulation is represented as a dot relating the achieved mean wait time across all requests and the measured empty distance covered by the simulated fleet. To guide the reader, all simulations performed with the same operational strategy are connected by black lines, while all cases simulated with the same fleet size are connected by blue lines. Following the black lines, hence, allows understanding how varying fleet size affects the operational characteristics of the system using a specific strategy while following the blue lines allows understanding how different operational policies affect the relevant metrics when operated at the same fleet size.

Finally, Figure 3 shows an area in which the real taxi system of San Francisco is supposed to operate. As we do not have information on the average taxi wait time from the data, all wait times are shared, while, for the empty distance, only the area above the optimistically estimated empty distance is covered.

4.2. Taxis in Chicago. A very extensive dataset of taxi trips is available for the city of Chicago. It is published by the city authorities and is available online (Chicago Data Portal). In comparison to the San Francisco taxi traces, no information about the paths of the taxis is given. Instead, only the trips are listed.

For each trip, information about its origin, destination, time, and payment is given. The set includes unique trip and taxi identifiers, a trip start time stamp in 15 min resolution, a trip duration in seconds, and the location of the centroids of the pickup and drop-off community areas.

The data is accessed through a web interface that allows specifying any time range in the past few years. Today, the data set consists of more than 187 million trips. As an example of a random day, we have picked July 19, 2019, for our analysis. Then, the scenario was created as follows: on July 19, a total of 57,876 trips appear in the data. A total of 5,910 cannot be processed because either their origin or their destination community area has been hidden by the publisher in order to guarantee the privacy of taxi users. A total of 2,402 trips were removed because they take less time than 2.5 min or longer than 3 hours. After these steps, 49,564 trips remain.

These remaining trips cannot be directly used for simulation for two reasons: first, their start time is only available in a 15 min resolution, that is, all trips' start times are rounded to 0, 15, 30, or 45 minutes of an hour. Second, the trip start and end locations are always at the centroids of the respective census tract or community area. The start coordinates are at 1 of 167 origins, and the end coordinates are at 1 of 251 destinations. In total, there are only 255 unique origin and destination points.

To have a scenario that resembles a real-world case more closely, the trip start times were randomly increased, drawing delays from a uniform distribution in $[0, 900]$ seconds. Then, all origins and destinations were resampled from a uniform distribution within their respective community area.

As an exception, the numerous trips at the O'Hare and Midway International Airports were kept as they are. Distribution within these community areas would have resulted in trips originating from the borders of the airport area, far away from the passenger pickup and drop-off zone. Clearly, this would not accurately represent real taxi trips.

The chosen approach of uniform spreading within the community area is conservative; a uniform distribution is

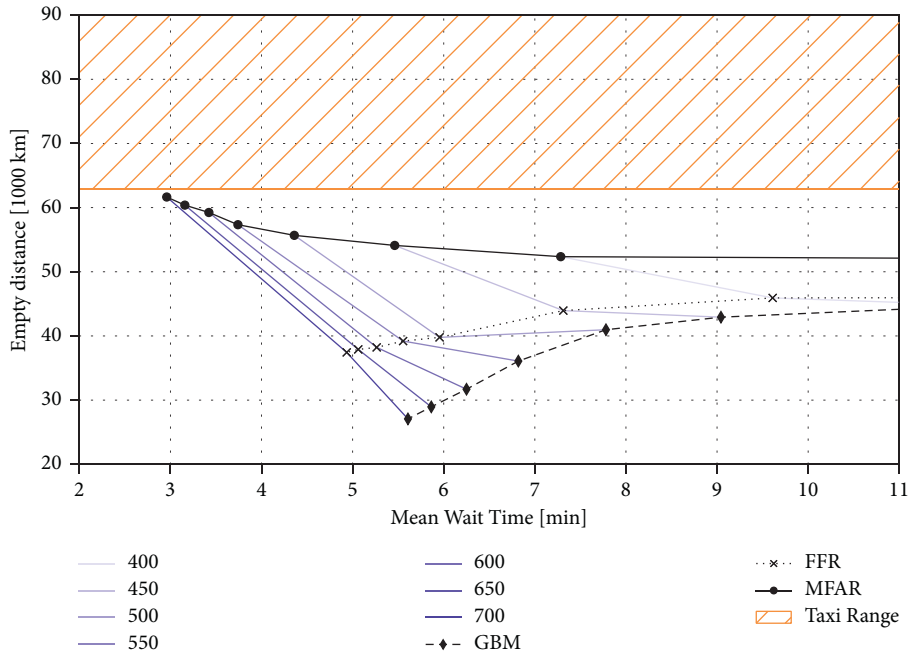

Figure 3: Mean wait time and empty vehicle distance of a coordinated mobility-on-demand system under different operational policies and fleet sizes for the San Francisco case. The real taxi system in orange uses 495 taxis. Wait times lower than in the coordinated system are unlikely but cannot be fully excluded based on the data.

more difficult to handle for mobility-on-demand systems than more pointed distributions, for example, [38]. The approach does not introduce unlikely accumulations of trips at certain locations. However, the approach does have the disadvantage that some of the modified trips could have unlikely combinations of origins, destinations, and recorded durations.

Thus, this modified trip set of 49,564 trips was filtered again. A total of 8,531 trips are faster than the network admits; 85 had a delay longer than 1 hour; 1,185 were slower than 5 mph; 196 were shorter than 200 m; and 353 started and ended on the same link. After this filtering step, a total of 39,786 sensible trips remained.

These trips are used to compute a network speed profile using Algorithm 1. Figure 4 shows the cumulative density function of both the simulated and the recorded trip durations. In this case, the fit is even better than in the San Francisco scenario and is assumed to represent an accurate model of road network conditions.

In total, there are 57,876 taxi trips on the chosen day, and a total of 3,950 unique taxis appear in these records. In the final scenario, there are 39,786 taxi trips and a total of 3,449 different recorded taxis.

If the number of taxis were reduced proportionally to the number of trips, we would expect 2,479 taxis, which is a more conservative estimate of the taxi fleet size and which is chosen for this reason. By computing the sum of intertrip shortest-time paths, we determine that these taxis served the final demand with at least 114,412 km empty distance. These values and the operational characteristics of the simulated coordinated mobility-on-demand configurations are shown in Figure 5.

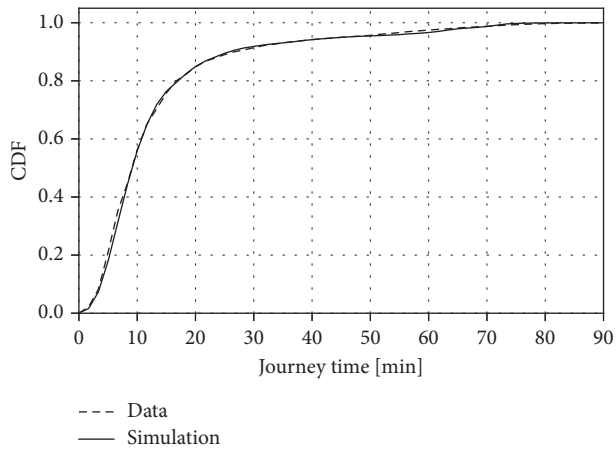

Figure 4: Comparison of journey times in simulation and in the original data set for Chicago taxis.

4.3. Taxis in Zurich. A taxi data set for the city of Zurich, Switzerland was provided to the research group by a local transportation company. It is different from the other used sets as it contains actual trip bookings made in a taxi call center. As in the other taxi data sets, the data allows extracting the origin, destination, duration, and start time of a taxi trip. Additionally, in this set also, the time of the request submission and the time of the taxi assignment are known. Thus, the data set allows computing the actual wait times experienced by customers. To protect the privacy of customers and business interests of the company, the data set cannot be made publicly available. However, we are able

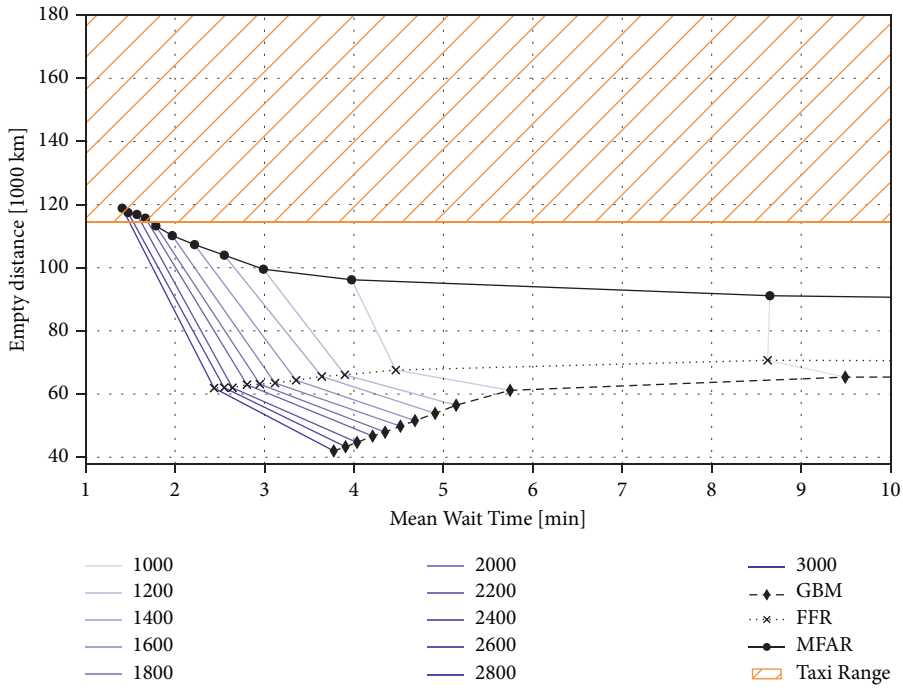

Figure 5: Mean wait time and empty distance of a coordinated mobility-on-demand system under different operational policies and fleet sizes for the Chicago case. The real taxi system in orange uses 2,479 taxis. Wait times lower than in the coordinated system are unlikely but cannot be fully excluded based on the data.

to present aggregate results in the work at hand and relate them to the other assessed taxi systems.

We have chosen June 21, 2017, for our analysis. On that day, the records contain a total of 1,248 trips that were booked in the call center. Four trips contain errors of some kind and are removed for this reason, for example, some of the required fields are empty. The remaining 1,244 trips were used for the simulation scenario.

Out of the 1,244 trips, 183 had travel times faster than the road network allows for and were thus removed for the estimation of the road network conditions. These remaining 1,061 trips were used to compute a network speed profile using Algorithm 1. Figure 6 shows the cumulative density function of both the simulated and the recorded trip duration.

As intended, the traffic conditions imposed on the simulated system are worse than the ones imposed on the actual taxi system. Overall, the curves have a good fit.

In the Zurich data set, the taxis belong to different operators or individual taxi owners that are independent of each other. The trips they get assigned to by the call center are only a subset of all their completed trips. Other trips, for example, curbside pickups, are not accessible. Thus, some taxis on record have only 1 or 2 trips during the entire day, although they have probably completed more trips.

This makes it harder to estimate the fleet size of the original taxi system for the 1,244 trips. As a conservative estimate, we thus assume that all taxis are as successful as the most productive taxi on record that served 13 trips assigned by the call center during the day. This allows conservatively estimating the fleet size for the taxi system at 95 taxis. By

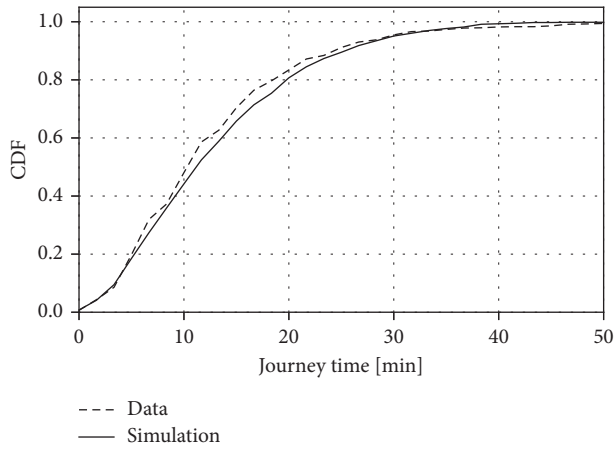

Figure 6: Comparison of journey times in simulation and in the original data set for Zurich taxis.

computing the sum of intertrip shortest-time paths, we determine that the taxi system exhibits at least 6,658 km empty distance to cover the trips in the recorded assignment and order.

Compared to the other data sets, the wait time every request experienced is known in this case: the mean wait time of requests in the real system is 6:46 minutes. Different from San Francisco and Chicago, the set of potential operational characteristics of the real taxi system in Zurich is represented as a line in Figure 7. It is located at the known wait time and covers distances above the optimistic lower bound of 6,658 km in terms of empty distance.

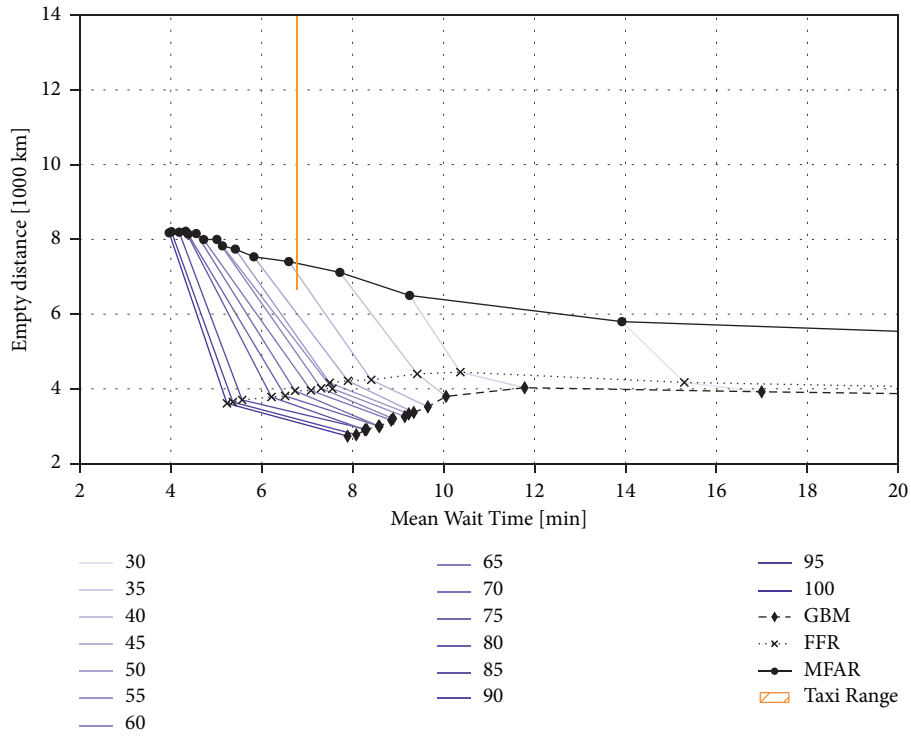

Figure 7: Mean wait time and empty vehicle distance of a coordinated mobility-on-demand system under different policies and fleet sizes for the Zurich case. The real taxi system in orange uses 95 taxis.

5. Discussion

The following sections discuss the results obtained above. First, the reference metrics obtained through our data analysis are summarized, followed by a per-city and per-algorithm analysis.

5.1. Baseline Metrics. Table 1 summarizes the estimated baseline situation for the three city scenarios. Both the trip count and the customer distance are directly derived from the taxi data. The analysis shows that the three cases cover a large range of differently sized demand cases from only 1,244 trips in Zurich to almost 50,000 trips in Chicago. Accordingly, we see customer distances at different scales. The estimated fleet sizes and lower-bound empty distances are summarized as well with fleet sizes from about 95 estimated vehicles in Zurich to almost 2,500 in Chicago.

The lower part of Table 1 shows derived measures that are interesting for analysis. The average trip distance is most accurate as it is computed from the data-based customer distance divided by the number of trips. The shortest average trip length is seen in Chicago with only 4.41 km in comparison to 6.65 km for San Francisco. Furthermore, Table 1 shows the trips per vehicle, the (estimated) empty distance per vehicle, and the total distance covered per vehicle (customer and empty distance). All three values are sensitive to the number of vehicles. A noteworthy value is a total distance covered per vehicle of about 370 km in San Francisco, compared to the substantially lower numbers in the other two cases. In fact, CRAWDAD states that the trips for San Francisco are identified by "cab number," while the trips in Chicago are identified by the "taxi medallion number" [45], which is linked directly to the driver. Hence, the values for San Francisco refer to the actual vehicle distance, while the data for Chicago (and Zurich) refer to driver shifts. UITP [46] shows that operating taxi vehicles in double shifts is a common practice worldwide. Dividing the trips per vehicle for San Francisco by two gives an average of 18.45 trips, which resembles strongly the value from Chicago.

Finally, Table 1 shows the share of empty distance compared to the total distance for all three scenarios. For San Francisco and Chicago, the value is close to 34% while it is almost 50% in Zurich. As stated in Section 4.3, the recorded trips used here are only a fraction of the actual trips performed by the drivers as only call-center requests are considered. It may be the case that the real empty distance ratio is substantially lower as drivers would perform additional trips between those visible to us. Hence, an analysis of absolute empty distance has been chosen in the previous sections in favor of the relative value. For comparison, UITP [46] lists a range of cities worldwide with estimates of empty distances from 30% to 50%.

In conclusion, the obtained reference values and their derived measures seem plausible and can be used as a basis for comparison with the simulation cases.

5.2. City Analysis. For San Francisco, the number of 495 vehicles has been estimated as the baseline fleet size. Looking at Figure 3, the closest evaluated case to that value is the graph for 500 vehicles (third point from the right for

Table 1: Summary of the baseline fleet characteristics derived from the taxi data in Sections 4.1–4.3.

	San Francisco	Chicago	Zurich
Data-based			
Trips	18,263	49,564	1,244
Customer distance (km)	121,400	218,581	6,727
Estimated			
Vehicles	495	2,479	95
Empty distance (km)	62,907	114,412	6,658
Derived			
Average trip distance (km)	6.65	4.41	5.41
Trips per vehicle	36.89	19.99	13.09
Empty distance per vehicle (km)	127.08	46.15	70.08
Total distance per vehicle (km)	**372.34**	134.33	140.89
Empty distance share (%)	34.13	34.36	**49.74**

MFAR). For all three operating policies, the optimistic threshold of the empty distance of about 62,000 km is not reached, that is, the system is more efficient. Depending on the strategy, a wait time from 4:21 min (MFAR) to 7:46 min (GBM) can be reached. At this fleet size, a reduction of empty distance from −11% (MFAR) to −34% (GBM) compared to the optimistic bound can be achieved. Real savings are expected to be even higher. The current optimistic empty distance is only reached with relatively large fleet sizes (here up to 700 vehicles) and if the waittime-optimizing MFAR strategy is used. The system could, hence, reach a wait time of only 2:51 min for the given requests through fleet coordination without becoming less efficient in terms of empty distance. On the other hand, the results show that a service operating with 700 vehicles and the GBM strategy could more than half the empty distance (−57%) at a waiting time of 5:36 min.

In Chicago, the analyzed configuration closest to the estimated 2,479 vehicles is the one with 2,400 vehicles (fourth graph from the left). Figure 5 shows that for both the FFR and GBM strategies, empty distance below the estimated optimistic threshold is achieved at this fleet size. For GBM, a reduction in empty distance of −60% can be obtained at a wait time of 4:12 min, while FFR provides another set point with a reduction of −45% and a wait time of only 2:48 min. The MFAR strategy at a fleet size of 2,400 vehicles operates at the optimistic empty distance with an average wait time of only 1:40 min that could be achieved.

Finally, the service operated in Zurich has an estimated fleet size of 95 vehicles, for which a simulation case is included in Figure 7. At this fleet size (second from left), the MFAR strategy exceeds the estimated empty distance of 6,658 km, while FFR and GBM stay below the threshold. The GBM strategy achieves a wait time of 8:40 min at a large reduction of the empty distance of −58%.

For Zurich, an additional analysis regarding the known average wait time of 6:46 min can be performed. When searching for those configurations that do neither deteriorate wait time nor exceed the optimistically estimated empty distance, all configurations in the sector left from the orange line (with lower wait times) and below its tip (with lower empty distance) can be considered. Note that Figure 7 shows that no set point for the MFAR strategy fulfills this criterion. Likewise, none of the analyzed cases for the GBM strategy adheres to these conditions. While Figure 7 does not strictly show that no viable configuration exists for GBM, it seems unlikely, given that the clustering of set points at the bottom of Figure 7 indicates that wait times saturate at around 7–8 min with increasing fleet sizes. Hence, only the FFR strategy provides relevant configurations in the Zurich case. To offer the same level of service as today, a minimum fleet size of 75 vehicles could be used with the FFR strategy. At the estimated fleet size of 95 vehicles, applying the FFR strategy in the running taxi system of Zurich may reduce the average wait time to 5:21 min at a reduction of the empty distance of −45%.

Similar waittime-based analyses could be performed for San Francisco and Chicago. Assuming that a wait time of five minutes would be acceptable for most users, different situations arise. For San Francisco (Figure 3), only one analyzed set point for the FFR strategy would fall into the viable range, while GBM would be excluded and a range of options using the MFAR strategy would be feasible. In contrast, for Chicago (Figure 5), most of the analyzed cases of all three strategies provide better wait times than 5 min and decreases in empty distance. It, hence, depends on the local context of the use cases that strategy should be applied to reach set goals in terms of customer experience (wait time) and efficiency (empty distance).

5.3. Algorithmic Analysis. In terms of algorithms, the expected effects of the three chosen algorithms can be observed: for all three cities, the GBM algorithm yields substantially less empty distance for fleet configurations that are chosen to operate at comparable wait times. On the contrary, the MFAR algorithm causes the largest empty distances. All three graphs show that when choosing the GBM, empty distance decreases with increasing fleet size. Theoretically, in the limit of an infinite number of vehicles, a threshold should be reached that can only be further pushed towards zero when considering the initial placement of vehicles. One could then, theoretically, reach the point where one vehicle is always available when a request appears. Note that no such guarantee exists for MFAR, which heuristically sends one additional empty vehicle to wherever a

request has appeared before. While zero empty distance for GBM could be achieved by having one correctly placed vehicle per request, the heuristic approach of MFAR would move these vehicles in a suboptimal way.

Regarding wait times, all three plots show a shift towards higher waiting times when focusing on particular fleet size and comparing MFAR and GBM. Hence, the MFAR algorithm systematically provides lower wait times at the expense of added distance. The FFR algorithm, finally, provides a trade-off between MFAR and GBM and becomes only the competitive choice among the three if both empty distance and wait time requirements are in place.

6. Conclusion

In this work, existing taxi systems were compared to hypothetical coordinated mobility-on-demand systems for different fleet sizes and operational policies.

The studies based on actual taxi data available for the cities of San Francisco, Chicago, and Zurich showed that fleet coordination is able to significantly improve both efficiency and service level of on-demand mobility. In order to reap these benefits, full control of the vehicle actions is necessary. In the current setup of many mobility-on-demand systems, for example, taxi or ride-hailing systems, this is not possible, as drivers take independent decisions and are also responsible for their own profit and loss.

Self-driving cars will inevitably overthrow this paradigm as all of their actions have to be determined by the operator, implicitly or explicitly. Thus, a change to an operational mode allowing for full coordination could be a worthwhile consideration for mobility-on-demand systems with human drivers, even ahead of the availability of fully self-driving cars.

Future research on this topic necessarily includes on-the-road case studies with real mobility-on-demand systems as well as the assessment of additional data sets, for example, for other cities or, if available, for ride-hailing schemes.

Data Availability

The data sources for taxi data of Chicago and San Francisco are named in the paper and accessible by anyone after registration. Additional data from OpenStreetMap are used. Data for the case study of Zurich are not publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the team of yourmile AG for their support in this project. The authors would like to thank their colleague Andreas Aumiller who provided insight and expertise that greatly assisted the research.

References

[1] F. Carbonero, E. Ernst, and E. Weber, "Robots worldwide: the impact of automation on employment and trade," ILO Research Department Working, 2018.

[2] H. Becker, F. Ciari, and K. W. Axhausen, "Comparing car-sharing schemes in Switzerland: user groups and usage patterns," *Transportation Research Part A: Policy and Practice*, vol. 97, pp. 17–29, 2017.

[3] A.-K. Hess and I. Schubert, "Functional perceptions, barriers, and demographics concerning e-cargo bike sharing in Switzerland," *Transportation Research Part D: Transport and Environment*, vol. 71, pp. 153–168, 2019.

[4] G. Dudley, D. Banister, and T. Schwanen, "The rise of Uber and regulating the disruptive innovator," *The Political Quarterly*, vol. 88, no. 3, pp. 492–499, 2017.

[5] M. Pavone, "Autonomous mobility-on-demand systems for future urban mobility," in *Autonomes Fahren*, pp. 399–416, Springer, 2015.

[6] The Telegraph, *A History of the new york Cab*, https://www.telegraph.co.uk/news/worldnews/northamerica/usa/8491507/A-history-of-the-New-York-cab.html, 2011.

[7] M. Daily, S. Medasani, R. Behringer, and M. Trivedi, "Self-driving cars," *Computer*, vol. 50, no. 12, pp. 18–23, 2017.

[8] C. Ruch, S. Richards, and E. Frazzoli, "The value of coordination in one-way mobility-on-demand systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1170–1181, 2019b, https://doi.org/10.1109/TNSE.2019.2912078.

[9] C. Ruch, S. Hörl, and E. Frazzoli, "AMoDeus, a simulation-based testbed for autonomous mobility-on-demand systems," in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3639–3644, IEEE, Maui, HI, USA, November 2018.

[10] D. Santani, R. K. Balan, and C. J. Woodard, "Spatio-temporal efficiency in a taxi dispatch system," in *Proceedings of the MobiSys 6th International Conference on Mobile Systems, Applications, and Services*, Breckenridge CO USA, June 2008.

[11] X. Dong, M. Zhang, S. Zhang, X. Shen, and B. Hu, "The analysis of urban taxi operation efficiency based on gps trajectory big data," *Physica A: Statistical Mechanics and Its Applications*, vol. 528, Article ID 121456, 2019.

[12] J. Cramer and A. B. Krueger, "Disruptive change in the taxi business: the case of Uber," *The American Economic Review*, vol. 106, no. 5, pp. 177–182, 2016.

[13] M. Pavone, K. Treleaven, and E. Frazzoli, "Fundamental performance limits and efficient polices for Transportation-On-Demand systems," in *Proceedings of the 49th IEEE Conf. on Decision and Control*, pp. 5622–5629, Atlanta, GA, USA, December 2010.

[14] C. F. Manski and J. D. Wright, "Nature of equilibrium in the market for taxi services," *Tech. rep.*vol. 619, pp. 11–15, 1967.

[15] G. R. Frechette, A. Lizzeri, and T. Salz, "Frictions in a competitive, regulated market: evidence from taxis," *The American Economic Review*, vol. 109, no. 8, pp. 2954–2992, 2018.

[16] G. W. Douglas, "Price regulation and optimal service standards: the taxicab industry," *Journal of Transport Economics and Policy*, pp. 116–127, 1972.

[17] H. Yang, M. Ye, W. H. Tang, and S. C. Wong, "Regulating taxi services in the presence of congestion externality," *Transportation Research Part A: Policy and Practice*, vol. 39, no. 1, pp. 17–40, 2005.

[18] E. C. Gallick and D. E. Sisk, "A reconsideration of taxi regulation," *JL Econ. & Org.*vol. 3, pp. 117–139, 1987.

[19] D. Cumming, "Why has the price of taxi medallions increased so dramatically? an analysis of the taxi medallion market," *The Park Place Economist*, vol. 17, no. 1, p. 9, 2009.

[20] L. Tang, F. Sun, Z. Kan, C. Ren, and L. Cheng, "Uncovering distribution patterns of high performance taxis from big trace data," *ISPRS International Journal of Geo-Information*, vol. 6, no. 5, p. 134, 2017.

[21] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: a recommender system for finding passengers and vacant taxis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2390–2403, 2012.

[22] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 899–908, ACM, July 2010.

[23] Y. Ding, S. Liu, J. Pu, and L. M. Ni, "Hunts: a trajectory recommendation system for effective and efficient hunting of taxi passengers,"vol. 1, pp. 107–116, in *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management*, vol. 1, IEEE, Milan, Italy, June 2013.

[24] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Online cruising mile reduction in large-scale taxicab networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 3122–3135, 2014.

[25] A. Horni, K. Nagel, and K. W. Axhausen, *The Multi-Agent Transport Simulation MATSim*, Ubiquity Press London, London UK, 2016.

[26] M. Maciejewski and J. Bischoff, "Large-scale microscopic simulation of taxi services," *Procedia Computer Science*, vol. 52, pp. 358–364, 2015.

[27] M. Maciejewski, J. Bischoff, and K. Nagel, "An assignment-based approach to efficient real-time city-scale taxi dispatching," *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 68–77, 2016a.

[28] M. Maciejewski, J. M. Salanova, J. Bischoff, and M. Estrada, "Large-scale microscopic simulation of taxi services. Berlin and Barcelona case studies," *Journal of Ambient Intelligence and Humanized Computing*, vol. 7, no. 3, pp. 385–393, 2016b.

[29] A. Poulhès and J. Berrada, "Single vehicle network versus dispatcher: user assignment in an agent-based model," *Transportmetrica: Transportation Science*, vol. 16, pp. 1–43, 2019.

[30] P. Jing, H. Hu, F. Zhan, Y. Chen, and Y. Shi, "Agent-based simulation of autonomous vehicles: a systematic literature review," *IEEE Access*, vol. 8, pp. 79089–79103, 2020.

[31] M. Hyland and H. S. Mahmassani, "Dynamic autonomous vehicle fleet operations: optimization-based strategies to assign AVs to immediate traveler demand requests," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 278–297, 2018.

[32] S. Hörl, C. Ruch, F. Becker, E. Frazzoli, and K. W. Axhausen, "Fleet operational policies for automated mobility: a simulation assessment for Zurich," *Transportation Research Part C: Emerging Technologies*, vol. 102, pp. 20–31, 2019.

[33] X. Zhan, X. Qian, and S. V. Ukkusuri, "A graph-based approach to measuring the efficiency of an urban taxi service system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 9, pp. 2479–2489, 2016.

[34] M. Haklay and P. Weber, "Openstreetmap: user-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[35] F. Poletti, P. Bösch, F. Ciari, and K. Axhausen, "Public transit route mapping for large-scale multimodal networks," *ISPRS International Journal of Geo-Information*, vol. 6, no. 9, p. 268, 2017.

[36] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research," *Proceedings - Institution of Civil Engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[37] M. Pavone, S. Smith, E. Frazzoli, and D. Rus, "Load balancing for Mobility-on-Demand systems," in *Proceedings of the Robotics: Science and Systems VII University of Southern California*, Los Angeles, CA, USA, June 2011.

[38] C. Ruch, J. Gächter, J. Hakenberg, and E. Frazzoli, "The +1 method: model-free adaptive repositioning policies for robotic multi-agent systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 3171–3184, 2019a, https://doi.org/10.1109/TNSE.2020.3017526.

[39] M. F. Hyland and H. S. Mahmassani, "Taxonomy of shared autonomous vehicle fleet management problems to inform future transportation mobility," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2653, no. 1, pp. 26–34, 2017.

[40] A. Mourad, J. Puchinger, and C. Chu, "A survey of models and algorithms for optimizing shared mobility," *Transportation Research Part B: Methodological*, vol. 123, pp. 323–346, 2019.

[41] J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, and D. Rus, "On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment," *Proceedings of the National Academy of Sciences*, vol. 114, no. 3, pp. 462–467, 2017.

[42] J. Bischoff, M. Maciejewski, and K. Nagel, "City-wide shared taxis: a simulation study in Berlin," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 275–280, IEEE, Yokohama Japan, October 2017.

[43] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "A parsimonious model of mobile partitioned networks with clustering," in *Proceedings of the 2009. First International Communication Systems and Networks and Workshops*, pp. 1–10, IEEE, Bangalore, India, January 2009.

[44] Crawdad, "A community resource for archiving wireless data," 2021, https://crawdad.org/epfl/mobility/20090224/.

[45] Chicago Data Portal, "Taxi trips," 2020, https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew.

[46] Uitp, *Global Taxi Benchmarking Study 2019. Tech. Rep.*UITP, Brussels, Belgium, 2020.