Recognizing Demand Patterns from Smart Card Data for Agent-Based Micro-simulation of Public Transport

Paul Bouman, Milan Lovric, Ting Li, Evelien van der Hurk, Leo Kroon, Peter Vervest
Rotterdam School of Management, Erasmus University
Burgemeester Oudlaan 50
Rotterdam, The Netherlands
{PBouman,MLovric,TLi,EHurk,LKroon,PVervest}@rsm.nl

ABSTRACT

In public transportation the question of how to achieve a good match between demand and capacity is essential for operators to provide a high quality service level within reasonable costs. Agent-based micro-simulation is a promising method to evaluate the impact of operational decisions and selected tariffs at both the level of the individual passenger and the aggregate level of the operator. During recent years, this technique has been applied successfully to several large scale real life cases. However, the demand of the agent population in these simulations is usually derived from aggregated census data and surveys conducted among a relatively small sample of the travelers. With the advent of smart card ticketing systems new opportunities to generate an agent population have surfaced. We use a unique smart card dataset containing four months of individual mobility data from passengers among three modalities in an urban Dutch public transportation system to generate agent populations. We model the temporal flexibility of agents based on patterns observed in the check-in/check-out behavior of individual travelers. We then run simulations to study how these agent populations react to a discounted tariff in the off-peak hours. Finally, we discuss opportunities to improve our approach in the future.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation, Algorithms, Management

Keywords

Agent Based Micro-simulation, MATSim, Pattern Based Demand, Public Transport, Revenue Management, Smart Card Data

1. INTRODUCTION

In public transportation systems without seat reservations, the question of how fluctuating demand can be serviced in a cost-efficient way poses a major challenge. Peaks in demand have a high toll on the costs, since they dictate the required amount of staff and the number of vehicles, while vehicles that are almost empty generate a net loss for the operator. Tools that allow the public transport operator to evaluate the effects of operational and strategic decisions on costs and demand are therefore vital to achieve the goal of improving the service quality and financial performance. However, most of the tools used in practice aggregate the passengers to homogeneous flows, either because detailed data is not available, or to reduce the complexity the decision maker has to face. During recent years, smart card systems have been introduced that log all movements of individual passengers through the systems. This gives a lot of detailed data that was previously unavailable. However, given the body of research related to smart card data, we can see that incorporating such data into the tools used for decision making is a non-trivial task [17].

A promising approach is agent-based micro-simulation. In such a simulation, individual passengers and vehicles are modeled through agents that interact with the public transportation system according to their individual goals. In this paper, we will use the MATSim simulation package [1] which has an active user-base and has been applied to a number of large scale scenarios. Within MATSim, all agents try to adapt their plans in such a way that their utility is improved. The simulation runs until there is no significant improvement within the agent population, i.e. until the population reaches an approximate equilibrium.

The major issue in generating an agent population from real life observations is the question how we can prevent agents to divert from this equilibrium in an unrealistic way, without restricting the agents in such a way that their only preference is to replicate the observed state.

We will limit our field of application to the study of *revenue management*. In revenue management[21] we want to control demand by adapting our pricing strategy in such a way that we get a better match between the available capacity and the demand emerging from the population. Our population can try to adapt to our pricing strategy by shifting the time at which they travel. We will study how the population reacts to an off-peak discount, but we believe that our approach is suitable for many other applications. One idea is to include the choice for mode of transport.

When generating our agent population, we run into the problem that the number of observed journeys differs a lot between individual passengers. We solve this problem by combining three types of demand that we can detect in our smart card dataset: *trip-based*, *tour-based* and *pattern-based* demand. Our first goal is to show how we can efficiently generate the agent population from our smart card data using these three demand models. Our next goal is to discuss how we can experiment with different parameters for the demand models to study revenue management. The final goal is to discuss our results and how we can improve our methods in the future.

The remainder of this paper is organized as follows: in Section 2 we discuss prior literature and related work. In Section 3 we discuss smart card datasets in general and our dataset in particular. Section 4 addresses the modeling of demand, based on the smart card dataset. In Section 5 we discuss the simulation and our experimental setup. We present the results of our experiments in Section 6. Finally, we discuss our results and opportunities for extensions of our approach in Section 7.

2. RELATED WORK

In recent years, smart card ticketing systems have attracted notable attention from the research community. A recent literature review on the use of smart card data in public transportation is given by [17]. They divide the studies into three categories: strategic-level studies, tactical-level studies and operational level studies. Since some of the public transportation systems only work with check-ins, part of the literature focuses on estimating the destination of passengers given their check-in location and time (for example, [23]). Some literature describes how the behavior of passengers can be analyzed. A notable example is [15], where spatial and temporal variations are measured across different types of cards. However, the literature review [17] contains not a single reference to the use of smart card data within a simulation context. Moreover, their conclusion contains the following quote:

> For the mass of data available on individual trips, new modeling methods will be needed, such as the Totally Disaggregate Approach, because classical models cannot be used at a such detailed level of resolution. [...] It will then be possible to calibrate individual base models from these large datasets. [17]

In the simulation of road traffic, microscopic simulation models have been a topic for quite some years. In the 1990's, it was mostly a topic studied as a field of application for super computers [10]. With the increase of computing power, more applications emerged in the 2000's, including [22]. With the introduction of MATSim [1], we saw a rise in literature related to micro-simulation. MATSim has been applied to some very large scale scenarios, including simulations of Berlin [19] and Zürich [13], both including more than a million individual travelers. Recently, MATSim was expanded from the simulation of road traffic, to the simulation of public transportation as well [18]. The website of the project contains a list with the most important publications related to the project and is updated regularly.

The kind of microscopic demand which is fundamental in the design of MATSim, is called *activity-based* demand [9] and was already discussed in the context of micro-simulation by [14] in 1997. This is an approach where travel demand is modeled by means of the activities the individual travelers want to perform over the day. One way to record the activities of individual travelers is by using surveys (for example [5]). In recent studies, census data was used to perform this synthesis of the activity based demand [4]. A survey on this approach to demand generation is given by [16].

Apart from modeling the activity patterns of travelers, a lot of research regarding the behavior of travelers has been performed, resulting in many sophisticated methods. Most notably, we would like to mention the field of discrete choice modeling [7], since it has spawned a lot of research within the domain of transportation. One of the main tools within discrete choice modeling is the stated-choice survey, where respondents have to select their preferred alternative.

A comprehensive textbook on revenue management is [21]. The focus of studies related to revenue management has been on systems where reservations are made in advance. In our setting, however, we do not have a mechanism where we can decide whether we accept new customers. This is different from, for example, long distance trains and the airline industry where tickets are always bought in advance. An example of a study related to revenue management in a comparable railway setting is [12]. This study shows some of the difficulties in applying revenue management within our context. An example of a succesful application for long distance trains with seat reservations is [8].

3. SMART CARD DATA

During recent years, the Dutch smart card, called "*OV-chipkaart*" was introduced as a cross-operator travel product. Starting from 2009, the smart card was made the mandatory product of travel in major Dutch cities, such as Amsterdam and Rotterdam, replacing paper tickets. One of the unique features of the Dutch system is that passengers have to check-in and check-out with the smart card in all modes of travel, including railways.

We use data collected from smart card usage over the course of four months from a major public transport operator in the Netherlands. During this period, the only avaible tickets were different smart card products. The transactions in our dataset denote either a check-in or check-out in a vehicle or on a platform. Moreover the smart card data contains the mode of travel, the unique id of the chip on the smart card (which we will call the media id), the time stamp of the transaction (in seconds) and the location of the transaction. Due to the sensitivity of the data for the operator and privacy concerns for the passengers, we will only show relative numbers and figures in this paper.

We prepared our raw dataset of almost 60 million transactions in such a way that we could process each transaction sequentially. We had to split up the dataset into separate chunks, using a round robin approach to assign media id's we had not seen before to a fixed chunk for that id. Afterwards, we sorted the separate chunks on media id and time stamp in main memory. We combined the results into a single dataset. While processing this set sequentially, we would be sure to encounter all transactions belonging to a certain media id together, with increasing time stamps.

After sorting the dataset, we linked check-ins and check-outs to make trips. Passengers who forget to check-out gives rise to inconsistencies in the dataset. It is relatively easy to filter these inconsistencies out, by assuming that a consecutive check-in and check-out belong together. This is reasonable, since the system has a maximum amount of time after which a check-in becomes invalid. After this linking step we know all the trips made by the passenger. Since the passengers have to check-in and check-out in each vehicle, we have separate trips when the passenger makes a transfer on his journey. Another preprocessing step is to link consecutive trips that are close in time to each other into journeys. This yields our main dataset. Figure 1a shows the numbers of unique passengers traveling over the course of a typical weekday. Figure 1b showsa histogram describing how many journeys were made with a single smart card. As we can see, most of the smart cards have made only a relatively low number of journeys, but there are plenty of passengers with many journeys.

4. DEMAND MODELING

When it comes to demand modeling for the simulation of public transport, a traditional approach is to use origin-destination matrices estimated from sources such as census data and manual counts of the number of passengers in some sampled vehicles [16]. The main drawback of this approach is that it becomes very difficult and expensive to measure the exact progression of passenger flows over the day. With smart card data, we know the origin, destination and exact time of travel of each individual travel, which allows for new opportunities with respect to measuring these flows.

Regarding flows of passengers in the network, we can take different approaches. The basic approach is to consider a flow through the network as a set of journeys: passengers who travel from a certain origin to a certain destination at a certain time. We will refer to this approach to demand as *trip-based* demand. However, in many cases there will be passengers who travel multiple times within the same day. In many of these cases, their consecutive journeys combine to a tour from origin to origin, with some intermediate destinations. In such cases, events happening at one of the intermediate destinations, will also influence the events in the remainder of the tour. Since our goal is to model individual passengers instead of aggregated flows, these tours contain valuable information. We will refer to this approach to demand as *tour-based* demand.

In activity-based micro-simulation, each individual traveler can be represented by an agent and this approach thus allows for microscopic analysis of a public transport system. The drawback is that we need a lot of information to model these agents. Even if we assume that all activities take place at a station, not all required information is available in the smart card data. The smart card data tells us *where*, *when* and *how* people travel, but it doesn't tell us *why* people travel, which is something that is vital to activity-based demand modeling.

Not all is lost, however: the traditional approach uses various statistical methods and interpolation techniques to fill the gaps of unknown information, in order to be able to simulate a public transport system. We can apply such an approach to the smart card data as well: we use the information which is available, such as location, modality and time of travel as much as possible and fill the gaps of information using estimation methods.

We will refer to the approach that goes beyond the notion of tour-based demand, but does not yet reach the precision of activity-based demand, as *pattern-based* demand. In a broad sense, we define pattern-based demand as demand produced by activities of such a nature that certain patterns will emerge in the travel behavior of passengers who perform the activity routinely. The most typical example of such an activity is working, since people usually work at regular times at a certain location. Other types of activities are education (which is usually bound to a schedule that may or may not change regularly), a periodic visit to family members and visiting sports events. In this paper, we will focus on patterns generated from working activities, since we believe that these will be most easy to recognize. In addition to this, we will consider educational activities with a fixed schedule as working activity, since the implications for the temporal flexibility of a passenger are usually similar. To summarize, we have:

Trip-based demand Demand with only a single journey.

Tour-based demand Demand consisting of a tour of journeys, with consecutive arrivals and departures at the same station. Also, the first and last station are equal.

Pattern-based demand Demand that exhibits a recurring pattern, produced by some regular underlying behavior of the passenger (which is possibly unkown).

4.1 Detecting customer patterns

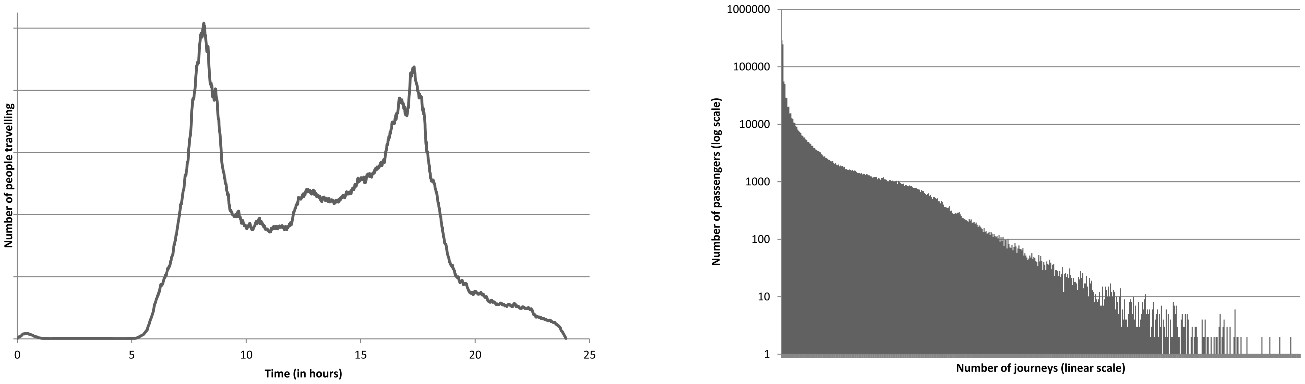
Commuters usually live and work at the same place. This leaves patterns of frequent *home* → *work* → *home* journeys in the smart card data. We can scan consecutive journeys for these patterns. This way we can derive an activity profile for a customer. For the sake of convenience, we limit ourselves to the class of activity profiles described in the following definition:

Definition 1. Activity Profile
An activity profile is a tuple $(l, b_{pref}, e_{pref}, \delta_b, \delta_e)$ where

- The activity takes place at location l
- The preferred starting time of the activity is b_{pref}
- The activity will not start before $b_{pref} - \delta_b$ and not after $b_{pref} + \delta_b$
- The preferred ending time of the activity is e_{pref}
- The activity will not end before $e_{pref} - \delta_e$ and not after $e_{pref} + \delta_e$
- The preferred duration of the activity is $e_{pref} - b_{pref}$

Now for each passenger, we will try to decide whether he is commuting and what his home and working stations are. To do this, we have to make a few assumptions.

1. We assume that somebody who is commuting travels a lot. Therefore, if the number of times traveled in the considered dataset is not above a certain threshold (which should be chosen according to the length of the time period under consideration), we conclude that the passenger is not a commuter.
2. We assume that a commuter has a fixed home and a fixed location of work and that the stations associated with these locations will be the two most frequently visited stations. To be sure these frequent stations are visited more frequently than other stations, we define thresholds for the number of times they should occur.
3. We assume that, if we include weekends, someone will spend more time at home than at work. Since we can measure the time between a consecutive arrival and a departure from a station, we classify the station where the greatest amount of time is spent as the home station.



(a) Demand histogram of a weekday

(b) Histogram of the number of passengers that made a certain number of journeys within 4 months

Figure 1: Demand as observed in the smart card dataset

4. We assume flexibility in time of travel and the length of the working activity is represented by a certain amount of variation in their travel times between their home and working stations.

We use the first assumption to decide whether we will try to recognize a pattern for a certain passenger at all. The second and third assumptions can be used to recognize a passenger's home station and working station. Finally, we use the fourth assumption to model the flexibility of a passenger based on this variance. These assumptions give us the following efficient algorithm:

Algorithm: Detecting Customer Patterns.

```
Parameters A minimum sample size $\theta$, thresholds $t_0$ and $t_1$ with $0 < t_0, t_1 \leq 1$
Input A set $J$ of $n$ journeys of a single passenger
Output A home station $s$ and a pattern $(t, b_{pref}, e_{pref}, \delta_b, \delta_e)$ that describes a working activity profile as defined in Definition 1
Step 1 if $n < \theta$ then conclude there is no valid pattern
Step 2 Find stations $a, b$ with maximal frequency as a start or endpoint over the journeys in $J$
Step 3 Denote $n_a, n_b$ as number of journeys that have $a$ or $b$ as a start or endpoint, $n$ as the total number of journeys in $J$
Step 4 if $\neg(n_a \geq t_0 n \wedge n_b \geq t_1 n)$ then conclude there is no valid pattern else
    Step 4a $\Delta_a :=$ average time difference between consecutive $(a, b)$ and $(b, a)$ journeys
    Step 4b $\Delta_b :=$ average time difference between consecutive $(b, a)$ and $(a, b)$ journeys
    Step 4c if $\Delta_a \geq \Delta_b$ then $s := a; t := b$ else $s := b; t := a$
Step 5 Take the average arrival time of $(s, t)$ journeys as preferred starting time $b_{pref}$
Step 6 Take the average departure time of $(t, s)$ journeys as preferred ending time $e_{pref}$
Step 7 Take the standard deviation of $(s, t)$ arrival times as the start time flexibility $\delta_b$
Step 8 Take the standard deviation of $(t, s)$ departure times as the ending time flexibility $\delta_e$
Step 9 return $s, (t, b_{pref}, e_{pref}, \delta_b, \delta_e)$
```

It is not difficult to see that each of the steps can be performed in time linear with respect to the set of journeys J, except for Step 2, where we have to calculate frequency statistics. To take the first and second most frequent station, we can sort the stations based on their frequencies. Since at most $O(n)$ station occur in J, this gives a $O(n \log n)$ time bound. In [20], it is discussed that this selection problem takes $O(n \log n)$ time in general. Since there are no loops in the algorithm, we may conclude that it runs in $O(n \log n)$ time for a single passenger with n journeys.

4.2 Deriving the Agent population

We will now discuss how to derive an agent population from our dataset. In the beginning of Section 4, we discussed the difference between *trip-based*, *tour-based* and *pattern-based* demand. Since there are smart cards that are used only once and passengers who have highly irregular travel patterns (because they don't use public transport to commute), we will not be able to derive a pattern for each customer and we may not even be able to find a tour in the data for each customer. Therefore, we will take a step-wise approach, where we first try to calculate a pattern for a passenger. If this succeeds, we will generate demand for this passenger based on the pattern we found. If we fail to find a pattern, we search for a tour and generate tour-based demand by introducing dummy activities at the intermediate stations of the tour. If we even fail to find a tour, we will generate trip-based demand by generating agents for each trip the customer made.

We will choose a single day (preferably not during the weekend) to model. We first filter our dataset such that we only retain customers that have traveled on that day. After filtering, we decompose our dataset into three parts: one group contains customers of which we know a lot, one group contains customers of which we have a tour and lastly, one group of customers with a single or unpredictable travel pattern. For each customer, we will have to generate an activity plan for the day. We will take a different approach to the generation of plans for each group of customers.

A plan for the day is a list of activity profiles with planned ending times for all activities. There is one exception: the last activity of the agent should be a home activity, which has no ending time. The ending time in the plan of an agent may differ from the ending times in the activity profile: an agent may try to deviate from his preferred time if this gives him an improvement in utility. The planned ending time is exactly what allows the agent to do this. When we start generating plans for our agent population, we will initially stick with the preferred ending times from the activity profiles as the planned ending times. For the group of customers for which we have derived a pattern, we can generate a *home* → *work* → *home* activity plan. For the group of customers for which we only have a tour, we only have a set of locations. For the activity profiles, we can easily derive a starting and ending time, using the check-out and check-in time at each intermediate station. The flexibility is a problem, however. For the time being, we decide to select a global value for the δ_b and δ_e of tour-based agents. We take a similar approach with the trip-based customers, where we generate a single agent for each trip. For each journey we observe from u to v, we generate an agent with a *home*→ *dummy*→ *home* pattern, where the first home activity should be performed at location u and the dummy and last home activity should be performed at location v. This gives us the following efficient algorithm for demand generation:

Algorithm: Generation of Demand.

```
Input A day $d$ and a set of customers $C$ with for each $c \in C$ their respective set of journeys $J_c$

Output An agent population for day $d$

Step 1 $P := \{p : p \in C, J_c$ contains a journey during day $d\}$

Step 2 $P_{pat} := \{p : p \in P, J_p$ has a pattern $\}$

Step 3 $P_{tour} := \{p : p \in P \backslash P_{pat}, J_p$ makes a tour at day $d\}$

Step 4 $P_{trip} := P \setminus (P_{pat} \cup P_{tour})$

Step 5 Initialize agent set $A := \emptyset$

Step 6 for each $p \in P_{pat}$
    Step 6a Generate an agent with a "home $\rightarrow$ work $\rightarrow$ home" plan
    Step 6b Add the agent to $A$

Step 7 for each $p \in P_{tour}$
    Step 7a Generate an agent with a plan containing the tour locations and ending times of $p$'s tour at day $d$
    Step 7b Add the agent to $A$

Step 8 for each $p \in P_{trip}$, for each $(u, v)$ journey traveled by $p$ on $d$
    Step 8a Generate an agent with a "home (at $u$) $\rightarrow$ dummy (at $v$) $\rightarrow$ home (at $v$)" plan of which the dummy activity should start at the check-out time of the journey
    Step 8b Add the agent to $A$

Step 9 return $A$
```

The running time of this algorithm is proportional to the size of the J_c sets. Let us define $n = \sum_{c \in C} |J_c|$. If we define $k = |C|$ as the number of customers and m as the maximum number of journeys for a single customer, we can easily see that $n \leq mk$. Steps 1-3 are regular filtering steps, that can be performed by examining each set J_c or by applying the earlier algorithm and can therefore all run in $O(mk \log m) = O(n \log m)$ time. The loops in steps 6-8 each iterate at most over k customers and generating the plan for each customer can be done in $O(m)$ time. Therefore, steps 6-8 run in $O(mk) = O(n)$ time as well. Therefore, the whole algorithm runs in $O(n \log m)$ time.

5. SIMULATION

5.1 MATSim

For our agent-based simulation, we used the MATSim 0.3.0 software package. To run a MATSim based simulation, we need three ingredients: the agent population, a network describing how vehicles can travel between nodes and a public transportation schedule. When we start the simulation, all agents calculate an initial plan. The main loop consists of a simulation and a replanning phase. During the replanning phase, each agent can adapt his activity plan. They do so by using certain modules available in MATSim, called mutators. During the simulation phase, all plans are executed and all events related to movements and activities of agents and vehicles are generated. The mutators used by the agents to adapt their plans, can be given individual probabilities. An example of such mutators are the rerouting mutator, that recalculates the fastest route between activities based on the network congestion of the previous day. Another example is the time mutator, that shifts the planned starting and ending times of the activities randomly, while retaining their sequential order.

Recently, the mobility simulation of MATSim has been extended with support for public transport [18]. This mobility simulation is an extension of the road-traffic simulation. In MATSim, model public transport vehicles as cars with a driver and a lot of space for additional passengers moving over a network that is given as input.

To generate the required network, we used a list of stations with their geographical locations and the available schedule information for all three modalities. We add the stations as nodes in the network. If there was a vehicle that visited two stations consequently in the schedule, we added a link between the two stations, with the distance of the link based on Euclidean distance between the two stops. We enforce the vehicles to wait at each stop until their scheduled time of departure. The mobility simulation itself is a discrete event simulation through a queuing network generated from the input network.

MATSim allows us to transfer money from or to an agent, but this mechanism is not triggered automatically. We added a module that imposes fares on the agents. It keeps track

θ	pattern	tour	trip
80	26%	32%	42%
120	14%	40%	46%
160	4%	48%	48%
200	1%	50%	49%
∞	0%	50%	50%

Table 1: Population distributions for different θ values

of the moments agents enter and exit the vehicles and the distances traveled by the vehicles. The fare of a journey consists of a *base tariff* that is the same for all journeys and a *distance tariff* with a certain fixed amount per meter traveled. An additional aspect is the transfer time: if the check-out time and check-in time of two consecutive journeys is small enough, the agent doesn't have to pay the *base tariff* a second time. At the end of the simulation of a single day, the accumulated fares are billed to the agent and transformed into disutilities during the evaluation of the executed plan. The utility function itself is described in [11]. The main idea is that traveling gives a disutility, while performing an activity gives utility. We did not yet implement an extension of this scoring function that assigns a personal price sensitivity to each agent, so this is currently a common parameter for all agents. We used 6 and -6 as the (global) coefficients for the performing and traveling utilities and -18 as the coefficient for late arrival.

5.2 Experimental Setup

We ran our experiments on a desktop PC with a quad-core Intel Core 2 Quad Q6600 processor and 8 GB of RAM running Windows 7 Professional SP1, 64-bit. Since we want our passengers to have their working station in at least half of their journeys and we want their home station to be at least as frequent as their working station, we chose $t_0 = 0.6$ and $t_1 = 0.5$. Prior to our experiments, we generated populations for different values of θ. We examined a few possible values, of which the distributions are presented in Table 1. Since $\theta = 80$, $\theta = 120$ and $\theta = \infty$ give us the greatest variations, we chose these for our experiments. For the tour-based and trip-based demand, we wanted our agents to keep as close as possible to their observed travel time, so we fixed δ_b and δ_b for their activity patterns to 5 minutes. This gives us a total of three different agent populations.

For our pricing strategy, we took figures inspired by the real world pricing policies. We set the base tariff to 0.75 and the distance based tariff to 0.115. The allowed transfer time is set to 30 minutes. For our experimentation with revenue management policies, we ran each of our populations through the network two times: once with a single tariff over the full day and once with a discount of 1% outside the peak hours (the peak hours are between 7:00–9:00 and 16:00–19:00). We chose 1% because our agents will always try to optimize their utility, even if the increase is very small. The check-in time determines whether the discount is given. To allow agents to shift their times, we enabled MATSim's time mutator module. Running each population against both pricing policies, we get a total of six experiments.

After some preliminary experiments, we saw that the increase of agent-utilities slowed down significantly between the 60th and 100th iteration. To be sure our simulation reached a state that is close to an equilibrium, we ran each simulation up until the 180th iteration.

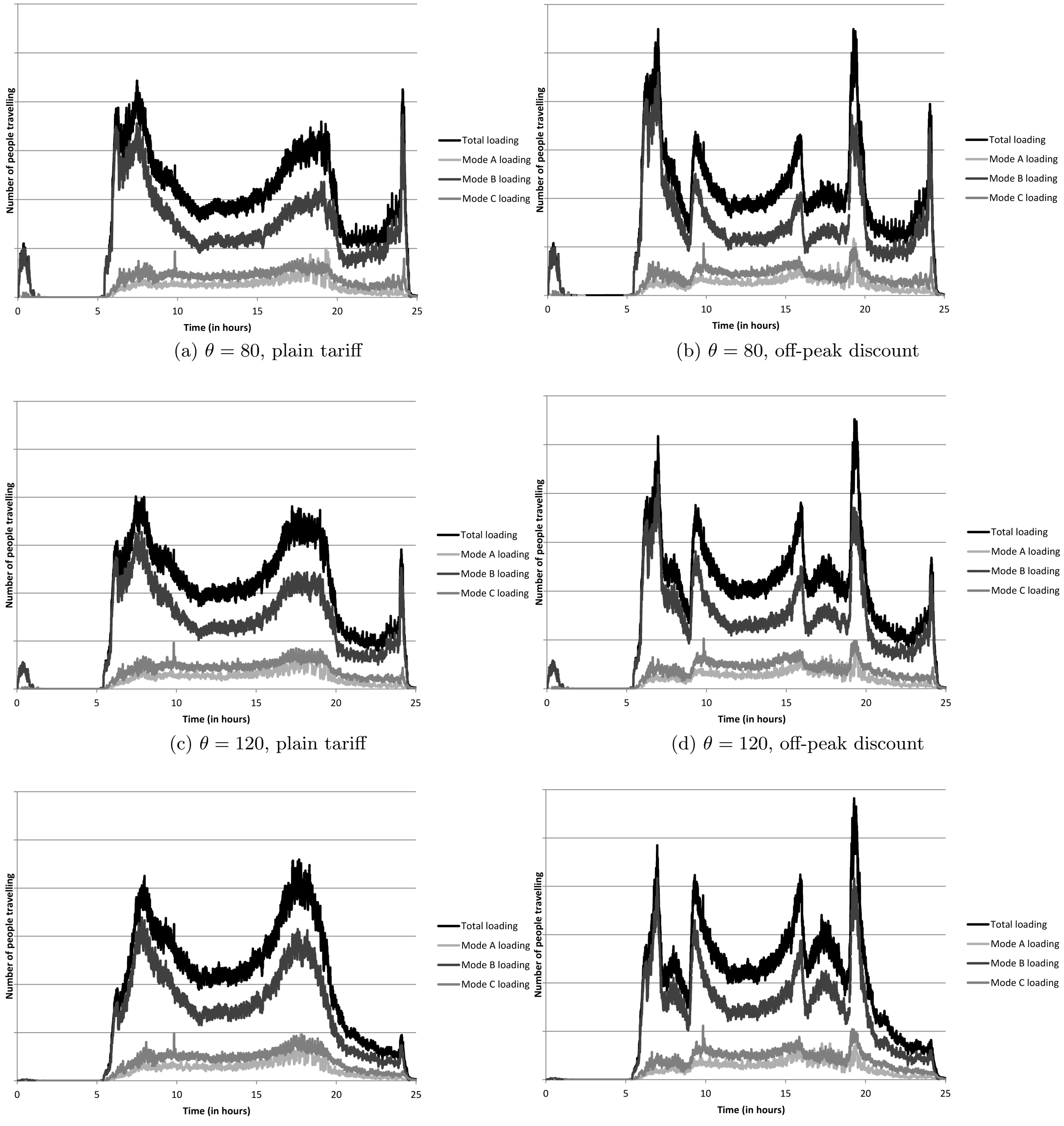
6. RESULTS

Generating our agent population could be done very efficiently from our sorted dataset of journeys that we derived in Section 3. The average time required to process this full set of 27 million journeys and write the agent population to MATSim input files was on average 107 seconds. Our simulation could roughly execute a complete iteration of the mobility simulation in a little less than two minutes. Some additional time was needed for finding all the shortest routes through the transit network and dumping all the plans of the agents after each 10th iteration. A complete run of a single scenario took roughly five hours. The vehicle loadings observed after the first iteration were in all of our six scenarios relatively similar to Figure 2e. But at the 180th iteration, we saw notable differences.

Let us first consider the case where we have a homogeneous pricing strategy over the whole day. When we move from $\theta = 80$ (Figure 2a) to $\theta = 120$ (Figure 2c), we see that the peak during the morning peak becomes a bit smaller, while the evening peak becomes a plateau that is a bit wider. This implies that, as soon as we treat some passengers who where pattern based in the $\theta = 80$ case as tour or trip-based during the $\theta = 120$ case, they tend to move away from the morning peak, but towards the evening peak. When we increase θ to ∞ (Figure 2e), we see that the morning peak increases a bit and the evening peak increases a lot. This suggests that some of the pattern-based agents in the $\theta = 120$ case actually traveled during the morning peak in the $\theta = \infty$ case, where they were less flexible.

One thing that should be noted is the high peak close to the end of the day in both Figure 2a and Figure 2b. This is a clipping artifact and implies that a certain group of agents prefers to travel at the end of the day and suggests there is a problem with the calibration of these agents. Although the problem decreases when we increase θ, the problem does not disappear entirely, even when we have $\theta = \infty$. We ignore this problem for the time being, but it suggests that we should be careful in drawing conclusions based on these results, and it is an issue that should be addressed in the future.

Now let us consider the scenarios where we discounted the off-peak hours. The most obvious result is the fact that this generates new peaks outside the peak-hour windows that are even higher than the rush hour peaks. This implies that even with a relatively small 1% discount, most of the agents have an incentive to divert from their initial plans. There can be two reasons for this behavior: either the agent is flexible enough to divert without losing utility, or the disutility of being early or late is smaller than the utility gained from the discount. We can study the result of decreasing the flexibility by comparing the results for $\theta = 80$ (Figure 2b) with the results for $\theta = 120$ (Figure 2d). A noticeable difference can be observed in the patterns that emerge within the peak-hour time windows. The evening peak in Figure 2d has a triangular structure, compared to the $\theta = 80$ case. When we increase θ to ∞, we get this triangular pattern in the morning peak as well and the effect is even stronger in the evening peak. Since all agents will travel by public transport and many agents diverted to the off-peak hours, the discount resulted in a drop in revenue.



(a) $\theta = 80$, plain tariff

(b) $\theta = 80$, off-peak discount

(c) $\theta = 120$, plain tariff

(d) $\theta = 120$, off-peak discount

(e) $\theta = \infty$, plain tariff

(f) $\theta = \infty$, off-peak discount

Figure 2: Vehicle loadings after 180 iterations for different sample size thresholds θ. On the horizontal axis, the time of day is displayed. On the vertical axis, the number of people currently travelling is displayed.

7. DISCUSSION

Our results show that our proposed method of generating an agent population from a smart card dataset and performing a microscopic simulation where each customer is presented by an agent is achievable within a reasonable amount of time. Generating the agent population and performing a single run of the simulation (given that all routes are calculated) both take under two minutes of time.

The results themselves show that the agents in our population react heavily to our discounted pricing policy, even if we have very inflexible agents in the $\theta = \infty$ case. However, we see that a certain number of agents still prefers to travel within the more expensive time window and in case of the $\theta = 120$ and $\theta = \infty$ cases, we see a triangular peak emerging within the peak hours. This suggests there is a population of agents for which the utility of arriving late is worse than the fare reduction. This should typically hold for agents who have to make a short distance trip, because for these agents the fare reduction is relatively low, compared to agents who have to travel a longer distance. This holds for real life passengers as well: a reduction on a small fare is of course much smaller than a reduction on a large fare. However, we can argue that the response of the agents is still too radical. We think the simulation will benefit greatly from calibration and utility functions that are not entirely linear with regard to the fare (especially when comparing prices, humans tend to disregard small price differences to some extent). Adding individual price sensitivities to our population of agents will be another way to improve in this regard.

When we compare the differences between our populations for different values of θ, we see that all populations maintain the property that during the typical peak hours demand is greatest. The value of θ seems to have the biggest impact on the evening peak. For lower values of θ, this part of the demand spreads outs to a much larger extent than the morning peak. This corresponds to the observation that usually the evening peak is longer in time and not as sharp as the morning peak. This is something which we can observe to some extent in Figure 1a as well. The main issue with the lower values of θ seems to be that we get greater clipping artifacts around 5:00 and 24:00. We hope this can be addressed by calibrating the utility functions, or by limiting the flexibility of agents when we come across individuals with extreme cases.

In the remainder of this section, we will discuss different topics for future research. In Section 7.1 we discuss how to improve the demand generation itself. We discuss the possibilities with regard to calibration of the parameters used by the simulation in Section 7.2. Section 7.3 addresses the question how we can incorporate additional datasets in order to distinguish different types of activities. Finally, Section 7.4 addresses the issue of validation.

7.1 Demand Generation

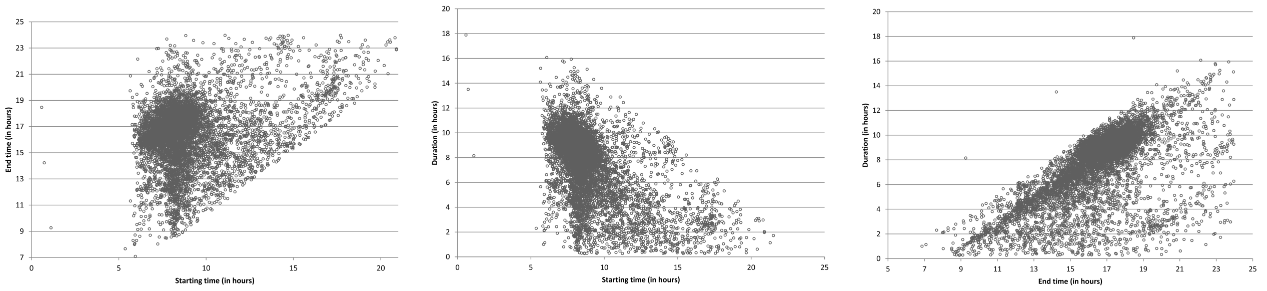
First of all, we must consider our demand generation algorithm. In our algorithm we make a couple of assumptions. The assumption that people who commute travel a lot, is very fundamental and probably realistic. The assumption that commuters have a fixed home and fixed working location probably often holds, but may be relaxed a bit: it can be broken by people who have more than one place to spend the night, or who have a job that has different locations that get visited in regular patterns. With enough observations, it may be possible to detect such patterns as well. The assumption that people spend more time at home probably holds often as well, but we must be careful with regard to outliers: it may be possible that somebody switches mode while at work (either by taking a bike or a car). In such an event, it would be possible that our approach reveals that somebody stayed for days at his working station, while this was not true in reality. The assumption that the variation in travel behavior of a passenger reflects his flexibility with regard to travel time is the most doubtful. Studies with more information regarding this assumption would be extremely valuable in improving our demand generation process.

While we have shown that our approach can efficiently generate an agent population from a real life smart card dataset, the fact that we have taken an approach that is very efficient and straightforward to implement has the disadvantage of being relatively crude. One may argue that we can introduce sophisticated pattern recognition and datamining techniques in this process, in order to generate an agent population that is closer to reality. One area for future improvement is that we use the average starting time and ending time of working activities, but ignore their possible correlations. In Figure 3 we can examine the scatter plot of the durations, starting and ending times of working activities performed by pattern-based individuals within the four months of our dataset. While a more thorough analysis is necessary, it seems probable that some correlations can be exploited. Improving our method in this regard is a priority, since we believe that this is useful information to make the behavior of the agent population more realistic.

7.2 Calibration

In order to reflect real life behavior more closely, calibration of our simulation is a required to use it in a decision support setting. The single global price elasticity for all agents is something that should be implemented on an individual level. We can do this in two ways. We can change the program to specify a utility function of each individual agent. Alternatively, we can adapt our fare-module to mimic price sensitivity. We can use a personal transformation function for each agent that scales the fares down for insensitive agents and scales the fares up for sensitive agents. Another kind of sensitivity that is valuable to model, is the sensitivity to the crowdedness of vehicles. If vehicles become too crowded, additional delay can induce delays in the public transportation system. This aspect was mostly ignored in our current simulation.

The right values for the price elasticities will be very difficult to estimate from only check-ins and check-outs. The main problem in this regard is the fact that we do not know what possible alternatives were available and have been considered by the passenger, before he made his journey. In the field of discrete choice modeling, this kind of data is referred to as *revealed choice* data. In situations where surveys are conducted and the subjects are exposed to multiple alternatives from which they must select a single option, we get *stated choice* data. Within the field of discrete choice modeling, most of the research effort has been performed on analyzing stated choice data. This allows us to accurately and efficiently estimate properties such as price elasticities within a population. In our case, it would be necessary to combine information obtained from stated choice experiments to calibrate the simulation obtained from revealed



(a) Starting time vs ending time

(b) Starting time vs duration

(c) End time vs duration

Figure 3: Correlations of starting times, ending times and duration ($\theta = 120$)

choice data. Some literature on how to combine revealed and stated choice data has been published during the 1990's ([6] and [3] are two examples). However, [17] reveals that there is little research in this area concerning smart card data.

7.3 Extensions based on additional datasets

One possible way to move our pattern based demand closer to real activity based demand models is by combining the smart card dataset with other datasets. A promising approach might be to look at regional information of stations. We could use such datasets to construct profiles of stations, which would allow us to make better guesses with regard to the activities that can be performed around the stations. If a station is close to a large industrial plant or office buildings, it is very probable that passengers traveling there do so because they have to work. A station close to a shopping mall will not only attract the employees of the shops, but customers as well. Local stations that coincide with a railway station or an airport are likely to attract passengers that want to travel further, or want to travel home. Stations in residential areas will likely serve as home stations, or as stations that get visited by passengers who want to visit friends or family. We propose to use data provided by the OpenStreetMap project [2], since this contains tags with information on available activities at certain locations.

After we generate profiles for all our stations with such information, we can take this information into account while recognizing patterns. This would allow us to make better guesses of the temporal flexibility of passengers for which we don't have a large enough set of journeys. Suppose we observe a passenger who starts his day with a journey from a residential area to an area with a lot of office buildings and stays there for 6 hours, then travels to an area with a shopping mall and stays there for 1 hour, after which he travels home. Even if we never observed any other journeys by this passenger, we can still make an educated guess about what he was doing and thus to what extend he could have been flexible. However, this calls for much more sophisticated statistical models than the one we are currently using. Depending on the kind of questions we want to study, it may or may not be worth the effort to go this far.

7.4 Validation

Validating a simulation like this is not a trivial task. One aspect that we can validate is the question whether the simulation can be used as a predictive tool for the movement of passengers through a public transportation network. The straightforward way to do this is by splitting the dataset at a certain moment in time. We can then use the first part of the dataset to generate agent populations and compare the outcomes to what is observed in the second part of the dataset. At first, we should choose a moment within a period where no policy and scheduling changes have occurred. If we can pass this test, we can raise the bar by choosing the moments at which a policy change has occurred, such as the introduction of a new schedule or new pricing schemes.

Another aspect that we may want to validate, is the question whether the emerging activity patterns of the agents represent the real-life activity patterns of the passengers represented by the agents. Validating this aspect requires much greater effort than validating the movements of passengers. One approach could be to use survey data containing activity logs registered in diaries and compare the diaries to the activity plans in the simulation. There may be some privacy issues with this approach, since it would require that we link the smart card id's to the participants, in order to match a diary to an agent. A possible workaround is to generate faux check-in/check-out data from the diaries by generating a check-in and a check-out for the journeys documented in the diaries. We could then use this dataset as if it were a smart card dataset and investigate to what extend the generated activity patterns of the agents reproduce the original activity plans.

In a similar way, we can consider the study of other location tracking datasets, such as triangulation logs from mobile phone operators or the location logs from the mobile phones themselves. The main advantage is that such a dataset contains more details on the whereabouts of individuals, which gives more opportunity to estimate what they are doing. For example, using smart card data we may observe that a person checks out at a station near a shopping mall and checks in four hours later. However, we have no data to decide whether it is probable that this person has been shopping or that this person has been working as an employee at one of the stores. If we have a mobile phone log, we may observe that the person has visited a great number of stores during these four hours. This would be evidence that he was not working as an employee.

8. CONCLUSIONS

We have shown how we can use smart card data to generate different types of demand. We developed an agent-based simulation that allows us to analyze the movements of the agents through our multimodal public transportation network. We experimented with different settings for the number of trip-based agents and with a 1% discount in the off-peak hours. Finally, we discussed several opportunities for future research.

As soon as we sorted our dataset in such a way that we could process all journeys customer by customer in chronological order, demand generation could be done very efficiently. We used simple rules to determine whether a customer should be modeled using *trip* based, *tour* based or *pattern* based demand. We have evaluated the impact of different thresholds for the *pattern* based customers on the resulting approximate equilibrium. We have also seen that an off-peak discount can be used to let a part of the agent population shift their travel times. In our case, this lead to a lower revenue. However, the effect on the required capacity must be taken into account when making a tradeoff between costs and revenue.

There are many opportunities for future research. First, our simulation can greatly benefit from proper calibration. Additionally, our method for demand generation can be improved upon, both by taking a closer look at the smart card data itself using more advanced techniques and by combining the smart card data with additional datasets. Including heterogeneity in the price sensitivity of the agents would be another improvement over the current situation. Finally, the simulation should be validated. We believe that an improved version of our simulation can be helpful in both the design of revenue management systems, including location based and modality based tariff schemes and other fields of study within a public transport context.

9. ACKNOWLEDGMENTS

We would like to thank Kai Nagel and his colleagues for their comments and support with the simulation. This research is sponsored by the Dutch National Science Foundation (NWO) Complexity Grant (#600.645.000.09). We also wish to thank Netherlands Railways for their support.

10. REFERENCES

[1] Multi-agent transport simulation toolkit, 2012. http://www.matsim.org.

[2] Openstreetmap, 2012. http://www.openstreetmap.org.

[3] W. Adamowicz, J. Louviere, and M. Williams. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of environmental economics and management*, 26(3):271–292, 1994.

[4] K. Axhausen, M. Balmer, and F. Ciari. A new mode choice model for a multi-agent transport simulation. In *8th Swiss Transport Research Conference. ETH, Eidgenössische Technische Hochschule*, 2008.

[5] K. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124, 2002.

[6] M. Ben-Akiva, M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. Combining revealed and stated preferences data. *Marketing Letters*, 5(4):335–349, 1994.

[7] M. Ben-Akiva and S. Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. The MIT press, 1985.

[8] N. Ben-Khedher, J. Kintanar, C. Queille, and W. Stripling. Schedule optimization at SNCF: From conception to day of departure. *Interfaces*, pages 6–23, 1998.

[9] J. Bowman and M. Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1–28, 2001.

[10] G. D. B. Cameron and G. I. D. Duncan. Paramics - parallel microscopic simulation of road traffic. *The Journal of Supercomputing*, 10:25–53, 1996. 10.1007/BF00128098.

[11] D. Charypar and K. Nagel. Generating complete all-day activity plans with genetic algorithms. *Transportation*, 32:369–397, 2005. 10.1007/s11116-004-8287-y.

[12] H. Link. Pep–a yield-management scheme for rail passenger fares in Germany. *Japan Railway & Transport Review (March)*, (38):50–55, 2004.

[13] K. Meister, M. Balmer, F. Ciari, A. Horni, M. Rieser, R. Waraich, and K. Axhausen. Large-scale agent-based travel demand optimization applied to Switzerland, including mode choice. In *12th World Conference on Transportation Research, Lisbon, July 2010*, 2010.

[14] E. Miller. Microsimulation and activity-based forecasting. In *Activity-Based Travel Forecasting Conference*, 1997.

[15] C. Morency, M. Trépanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[16] K. Müller and K. Axhausen. *Population synthesis for microsimulation: State of the art.* ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT), 2010.

[17] M. P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557 – 568, 2011.

[18] M. Rieser. *Adding transit to an agent-based transportation simulation.* PhD thesis, Technical University Berlin, Berlin, 2010.

[19] C. Rommel. Automatic feedback control applied to microscopically simulated traffic the potential of route guidance in the berlin traffic network. Technical report, VSP Working Paper, 2007.

[20] S. S. Skiena. *The Algorithm Design Manual.* Springer, 2nd edition, Aug. 2008.

[21] K. Talluri and G. Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Verlag, 2005.

[22] M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805, 2000.

[23] M. Trépanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.