Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

REAL-TIME PERFORMANCE MEASUREMENT SYSTEM FOR AUTOMATED TELLER MACHINES

R.G. van Anholt

VU University Amsterdam
Faculty of Economics and Business Administration
Boelelaan 1105
Amsterdam, NL-1081HV, THE NETHERLANDS

I.F.A. Vis

University of Groningen
Faculty of Economics and Business
Nettelbosje 2
Groningen, NL-9747AE, THE NETHERLANDS

ABSTRACT

Performance measurement systems have proven to facilitate process improvement in the past decades in various markets and environments. The objective of this paper is to design a performance measurement system to actively control, monitor and improve performance of automated teller machines. In our real-time performance measurement system we apply different weights for different types of potential lost sales. We implement the measurement system in an ATM inventory management context and conduct discrete event simulation experiments to demonstrate its added value in terms of cost and service. We use data from Dutch commercial banks in our performance and sensitivity analyses. Exhaustive numerical validation demonstrates that the implementation of our performance measurement system leads to a higher fill rate – 99% instead of 98% – with equal expenses, or a 3.7% cost reduction while maintaining an equal fill rate.

1 INTRODUCTION

For centuries, paper money (i.e., cash) has been the foremost payment instrument worldwide. Only in the past century, cash received competition from mainly checks, credit and debit cards. Nonetheless, in the United States of America and the Eurozone, 80% of all 895 billion transactions were still made using cash in 2011. Because a further reduction in the usage of cash is expected (G4S Cash Solutions 2011, Capgemini, RBS, and EFMA 2011), banks try to improve the processes of cash supply chains to reduce the cost per processed banknote. Significant supply chain cost is incurred by Automated Teller Machine (ATM) related activities. These ATM activities are therefore common key issues for process improvement. In this paper, we perform exhaustive numerical experiments by means of a discrete event simulation model to show that improvements in ATM performance measurement reduce ATM manager's cost and/or increase the perceived service quality by ATM users.

In general, a cash supply chain consists of a central bank (responsible for cash quality), cash centers (responsible for distributing, checking, sorting and preparing cash), bank branches and ATMs (which dispense and take cash to/from customers), retailers (which primarily receive cash from customers), and customers (who are the end-users of cash). The ATM manager, either a commercial bank or an individual ATM deployer, is responsible for ATM activities such as maintenance and inventory control.

Striving for economies of scale is one solution to reduce supply chain cost. In The Netherlands, for example, the three largest banks have recently decided to transfer the cash processing, distribution, and ATM maintenance gradually to a single company to obtain economies of scale. Process improvement can be achieved by implementing a performance measurement system (PMS) as well. A PMS consists of multiple interrelated performance metrics which can be used to monitor, measure and control ATM service quality. Most research in PMSs focuses on integrating performance metrics into manufacturing systems.

The main thought is that a dynamic set of metrics relevant to each managerial level has to be defined instead of using individual performance measures (Ghalayini, Noble, and Crowe 1997). Recently, the added value of deploying multiple PMSs within a supply chain has been studied (Cagnazzo, Taticchi, and Brun 2010). The authors demonstrate that having multiple PMSs within the supply chain is a critical success factor for the entire supply chain .

To our knowledge no PMS for managing ATM replenishments and maintenance is available in literature. Practical studies, however, demonstrate the need. In collecting empirical data at commercial banks in The Netherlands, we noted that banks struggle with the exact definition of performance (Van Anholt and Vis 2010). Consensus exists among commercial banks on the importance of two key service quality dimensions; cost and service, but the actual meaning and usage varies substantially between commercial banks. Every bank aims for low operational cost and high perceived customer service quality, but each bank measures performance dimensions in a slightly different way. The lack of a PMS that covers best practices has been the motivation for conducting the research presented in this paper.

Generally, ATM cash managers have to comply with service level agreements (SLAs) which involve a target fill rate that needs to be fulfilled as an average for a network of ATMs within a time interval. A fill rate is often referred to as the fraction of the demand filled from on hand inventory (Kleijnen and Smits 2003). Demand uncertainty, lead-time uncertainty, and ATM breakdowns keep the fill rate from reaching 100%. We cast doubt on whether this definition of a fill rate is the appropriate metric to measure service quality. For example, users would mind an unavailable ATM more if this ATM is the only ATM in walking distance, than when the unavailable ATM has another (operational) ATM located next to it. We argue that the impact of not fulfilling customer demand on customer satisfaction depends on ATM characteristics and the moment of unavailability. Customer satisfaction is the extent to which the results produced for the customer and the process he or she went through to secure the results meet with his or her expectations (Harvey 1998). We demonstrated that the degree of customer (dis)satisfaction highly depends external factors by conducting a survey with 2,950 Dutch ATM users (Van Anholt and Vis 2012).

To our knowledge, PMSs in both literature and practice do not differentiate between weights of missed transactions, i.e., each missed transaction is considered to impact customer satisfaction equally. The real-time PMS we develop considers differences in weights of missed demand and can be utilized for managing all ATM activities concerned with delivering ATM service quality. The real-time aspect is important because the weight of missed demand is, among other things, a function of time. Next to designing the PMS in this paper, we also add value by demonstrating in a simulation study that implementing the PMS into the inventory management of ATMs improves fill rates and/or efficiency. Simulation allows us to easily model the real-world stochastic variables and complex decisions in a dynamic setting.

Among related literature, Vasumathi and Dhanavantan study the ATM queuing problem using simulation. The developed simulation technique reduces the idle time of ATMs as well as the waiting time of users for any commercial bank with ATM services. Other researchers (Suri, Singh, and Chander 2007) use simulation to determine efficient locations for future ATMs. Also the inventory routing problem of ATM networks has been studied (Wagner 2010). Wagner propose simulation and factorial design methods to analyze the integrated problem of inventory management and vehicle routing. To our knowledge, discrete event simulation has not been used for analyzing ATM service quality before.

In Section 2 we present our real-time PMS. Section 3 elaborates on the key decisions in ATM inventory management and how the real-time PMS interacts with these decisions. A description of our discrete event simulation study is presented in Section 4. Results are included in Section 5, followed by the sensitivity analysis in Section 6. Section 7 presents conclusions.

2 REAL-TIME PERFORMANCE MEASUREMENT SYSTEM

ATM unavailability has several causes; most frequently an ATM is out of service due to a (technical) malfunction, an ATM might also be down because it ran out of banknotes, and a third and last reason of unavailability might be due to an ATM replenishment. ATM users do not leave a trace of visiting an ATM when it is unavailable, so the exact amount of missed demand during ATM unavailability is un-

known. Consequently, only the downtime can be measured. ATM downtime as sole performance metric is insufficient because it ignores the huge variation in demand volume (see Figure 1 for an example). Different demand volumes can be expected; for example, more demand during daytime than at night, and also more demand during Saturday than during Sunday.

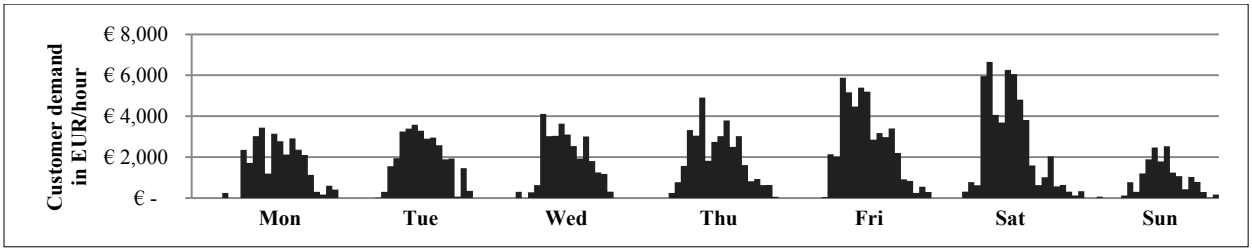

Figure 1: A single ATM's real hourly demand in Euro during a single week

We conclude that an estimate of the amount of missed demand for each coming day and hour is required to calculate the regular fill rate (see (1)). An estimate of the amount of missed demand can be derived from the demand forecast. When ATM inventory management is performed properly, a reliable demand forecast is readily available. If not available, the demand forecast needs to be calculated first.

$$regular\ fill\ rate = 1 - \left(\frac{estimated\ missed\ demand}{total\ demand\ \ fulfilled + estimated\ missed\ demand} \right) \tag{1}$$

Both the ATM characteristics and the moment of ATM unavailability determine the degree of customer dissatisfaction. For example, during a holiday or event an ATM user would be more disappointed when encountering an unavailable ATM than during a regular day. Commonly, a fill rate (see (1)) presumes that every missed transaction has an equal impact on ATM user's satisfaction and is therefore inadequate. In order to apply different weights for different types of missed demand, we consider an adjusted fill rate as additional performance metric. The adjusted fill rate (see (2)) presumes that each type of missed demand has a unique weight.

$$adj.\ fill\ rate = 1 - \left(\frac{weighted\ estimated\ missed\ demand}{weighted\ total\ demand\ \ fulfilled + weighted\ est.\ missed\ demand} \right) \tag{2}$$

To determine the weight of each type of missed demand we refer to a list of influential factors and elements (see Table 1) which is derived from Van Anholt and Vis (2012). According to commercial banks, individual ATM deployers and ATM users, these factors and elements have a moderating impact on the weight of missed demand. For example, users would be more disappointed if an ATM is unavailable where users can choose a denomination mix (i.e., composition of banknotes) themselves, than when an ATM is unavailable which does not offer this service. Next to assigning a higher weight because of service related arguments, a higher weight might also be preferred for ATMs that are profitable. ATMs can be profitable when users have to pay a surcharge for withdrawing cash. If surcharge fees apply, missed demand can be translated directly into lost income. The moderating effect of each element in Table 1 highly depends on the economic and demographic profile of the local market. Therefore we cannot present generally applicable moderating effects for each element in this paper. For now, we assume that the weight of missed demand for a certain ATM at a certain time can be calculated using the factors and elements in Table 1.

Table 1: Factors and elements increasing the weight of missed demand

Factors	Elements
1. Moment of ATM usage	▪ during holidays ▪ during (local) events
2. ATM characteristics	▪ if the ATM has a deposit function ▪ if the denomination mix can be chosen ▪ if the user can check his/her account balance
3. Location	▪ if not located next to other ATMs in walking distance ▪ if the ATM is located in a bank branch ▪ if the ATM generally has a long queue
4. Profitability	▪ when the ATM is profitable due to surcharge fees

The adjusted fill rate will not pay off if no changes are implemented in decision making. We need to avoid missed demand with a high weight, because this would decrease the adjusted fill rate more quickly. Hence, we should concern less about missed demand with a low weight because this would decrease the adjusted fill rate less quickly.

ATM managers need to comply with a certain target fill rate. The fill rate and other metrics are monitored during a fixed time interval; for example, a month, a quarter or a year. At the end of each time interval, performance metrics are registered and reset to zero. Hence, when the time interval runs to an end, the performance objectives should be met. For example, a target fill rate of 98% means that the realized fill rate should be equal or higher than 98% at the end of the time interval. Real-time monitoring of the realized fill rate is crucial because the amount of already missed demand determines how much demand can still be missed until the end of the time interval to achieve the target.

In Figure 2 a performance dashboard is included to show how the adjusted fill rate can be monitored in real-time. The performance dashboard can be constructed by keeping track of the amount, the cause and the weight of the estimated missed demand. The bar chart depicts the estimated missed demand in percentages, which coincide with the regular fill rate. We distinguish between a range of weights of [0.1 - 1.9] and we colored missed demand with a higher weight red and missed demand with a lower weight green. The adjusted fill rates have slightly different values than the regular fill rate because they are obtained by multiplying the missed demand with the respective weights. In the example in Figure 2 we observe that most demand has been lost because of ATM failures. We also see that each type of unavailability led to missing demand with both higher (colored red) and lower (colored green) weights. In practice, it could occur that the missed demand with higher weights (or lower weights) are primarily generated by one type of cause only.

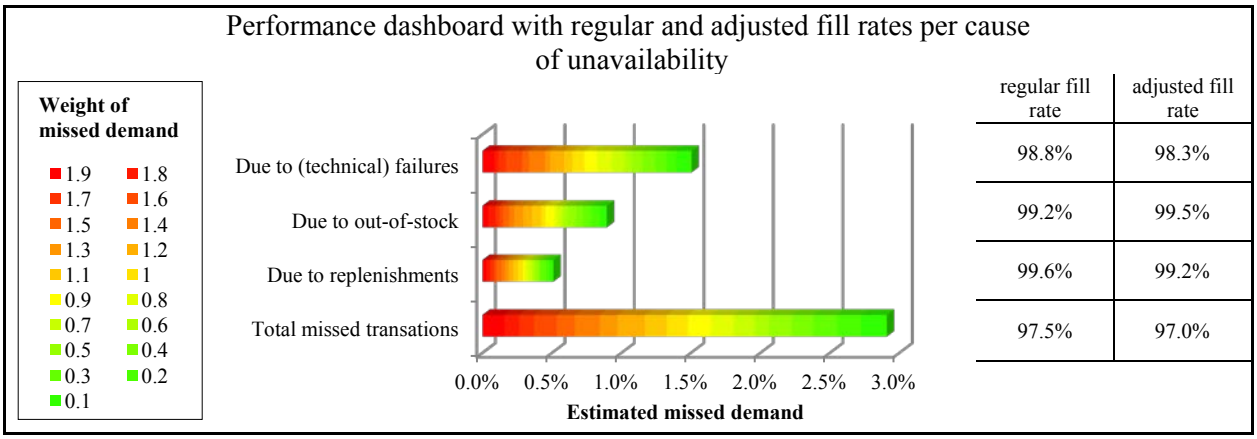

Figure 2: Performance dashboard of the regular and adjusted fill rates with numerical data

In addition to the adjusted fill rate, the holding and ordering costs are important metrics in the real-time PMS as well. The holding cost is also denoted as the interest cost or opportunity cost of cash, because cash owners do not receive interest income from cash that is stocked in ATMs. The ordering cost consists of the order preparation cost at a cash center, the transport cost, insurance cost and the cost of the actual replenishment. Transportation is particularly expensive because cash is a high value, high density product; transportation requires specialized armored trucks, delivery by at least two persons and insurance. The following section describes how the real-time PMS can be integrated in an ATM inventory management context.

3 REAL-TIME PMS AND ATM INVENTORY MANAGEMENT

This section addresses the integration of our real-time PMS in an ATM inventory management context. It is important to not only measure the adjusted fill rate, but also act accordingly. We discuss how decisions need to be taken in an ATM inventory management context while continuously using information provided by the real-time PMS.

Inventory management of ATMs is concerned with several interrelated key decisions. We consider the inventory management problem of choosing efficient reorder moments and efficient order quantities of ATM replenishments such that a target fill rate is achieved. When an ATM replenishment is ordered by an ATM manager, the order is consecutively prepared in a cash center, cross-docked at an armored carrier's vault, transported to the ATM and finally the ATM is restocked with the ordered cash.

An ATM's cash hold usually consists of four cassettes, each capable of storing up to two thousand banknotes of the same denomination. Given the currency denomination and customer demand of the respective market, ATM managers decide about the capacity of ATMs. In The Netherlands, for example, most ATMs are fitted with 2x€50, 1x€20 and 1x€10, which adds up to a capacity of €260,000. Due to customer demand, inventory decreases over time and replenishments are required. A lead-time between ordering and delivery needs to be taken into account. A demand forecast is required to estimate whether enough inventory is available to cover the order lead-time and to decide when an order needs to be placed. On every workday there is a specific cut-off moment (e.g., 11.00AM) before which a replenishment order can be placed. If an order is placed after this moment, the delivery will be performed after one additional workday. Due to ordering cost it is wise to postpone ordering as much as possible. Roughly, an order will be placed only if postponement would lead to a significant amount of missed demand.

The order quantity mainly depends on the holding cost of cash. To reduce the holding cost, small order quantities are preferred. Small order quantities however, result in frequent ordering which is expensive as well. Just like in traditional inventory problems, ordering cost and holding cost are continuously in debate. In principle, we apply the Silver-Meal heuristic (Silver and Meal 1973) to calculate the order quantity, which involves determining the order quantity to fulfill a number of review periods of future demand which results in lowest overall cost. We slightly adjust this heuristic to account for ATM capacity constraints, non-delivery days, variable lead-time and stochastic demand.

We fully integrate the real-time PMS in an inventory replenishment model that is purposely designed to perform ATM cash management (Van Anholt and Vis 2010). This inventory replenishment model considers a periodic review policy, non-stationary stochastic customer demand (with irregularities such as holidays, events and disturbances like ATM breakdowns), a stochastic replenishment lead-time, a target fill-rate and an ATM capacity restriction. We also assume specific delivery weekdays (e.g., not during the weekend), a fixed holding cost per stocked unit of cash and a fixed ordering cost for each replenishment. The inventory problem for ATMs is different from similar problems in other areas because of the combination of; relatively high holding and transport costs (due to safety and insurance), a homogeneous product, certain demand characteristics (difficult to forecast accurately because of high variability, seasonality and trends), certain lead-time characteristics (unknown delivery time and specific delivery working days and hours), and a target fill rate to comply with.

We elaborate in the remainder of this section on the decision of when to order an ATM replenishment. The decision of when to order is basically choosing between two options upon each reorder moment; order 'now' or postpone ordering (see Figure 3).

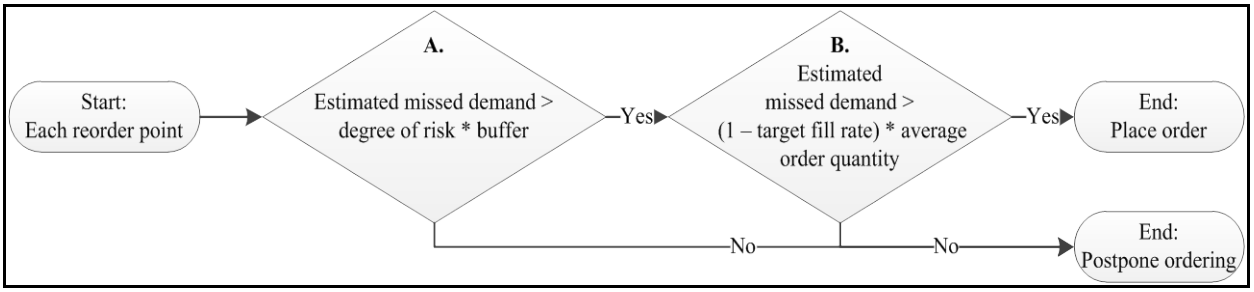

Figure 3: Schematic overview of the decision "when to order a replenishment"

A target fill rate (by definition smaller than 100%) presumes some percentage of demand is allowed to be missed. We estimate the amount of demand that can be missed until the end of the measurement time interval such that the target fill rate can still be achieved. We refer to this amount as the 'buffer'. As the name suggests, the buffer is initiated to cope with ATM unavailability, caused by uncertainty in demand, lead-times and breakdowns. To ensure that efficiency is obtained, the buffer should be used gradually during this interval and the buffer should be (near to) depleted by the end of the measurement time interval. The entire buffer should not be blown entirely at the beginning of the measurement time interval to avoid having high safety stocks later on. Also, being too careful and not using the buffer (enough) would incur high cost as well. Therefore, at each reorder point, ordering is only postponed if this would not lead into allocating more than a specific percentage (referred to as the 'degree of risk') of the buffer (see decision A in Figure 3). The input parameter 'degree of risk' is the fraction of the buffer that can be allocated at each reorder point. The effect of choosing different values for the 'degree of risk' on performance is checked in the sensitivity analysis in Section 6.

If the buffer is really small or negative, decision A would lead in extremely high safety stocks. Therefore one exception has been included: Postponement is also allowed when the expected amount of missed demand is smaller than (*1–target fill rate*) * *average order quantity* (see decision B in Figure 3). Decision B basically ensures that the achieved fill rate stays within the target fill rate or ensures that the achieved fill rate is not getting worse.

To achieve the *adjusted* fill rate we integrate the real-time PMS into the inventory management system by calculating the buffer with the weighted amount of missed demand. The integration ensures that the buffer decreases only slightly when demand with a low weight is missed and the buffer decreases substantially when demand with a high weight is missed. More costs will be incurred for avoiding stock-outs when missed demand has a high weight and less costs will be incurred when potentially missed demand has a low weight. We state that overall inventory management cost reduces or customer satisfaction improves by adopting the real-time PMS. We construct the following hypotheses:

Hypothesis 1 Integrating our real-time PMS in an ATM inventory management context reduces total ordering and holding cost while the same fill rate is achieved.

Hypothesis 2 Integrating our real-time PMS in an ATM inventory management context increases the realized fill rate while the same cost level is maintained.

In the next section a simulation model is proposed in order to test both hypotheses in Section 5.

4 SIMULATION MODEL

A mathematical programming based approach is difficult to realize because of the uncertainty in both customer demand and replenishment lead-time. Also, the target fill rate as input parameter is difficult to deal with in an optimal approach. Considering this, a discrete event simulation (DES) based approach is chosen. DES allows for easy modeling of dynamic and continuous systems (Law and Kelton 2000). We design a DES model of ATM replenishment decisions in which the real-time PMS is integrated.

Important to conducting a simulation study is verification and validation. We used the verification and validation paradigm of Sargent (2005). While building the model we used animation and kept track of variables to check whether the functioning of the model was credible, that is, if decision variables, tallies and output parameters show plausible values. To check the face validity; the input parameters, the decision variables and the behavior of the model are discussed with cash managers from Dutch commercial banks. We controlled for the internal validity by performing multiple simulation replications to determine the stochastic variability in the output parameters. A reasonable amount of variation was found, but not too much to question the appropriateness of the system. The event validity is examined in Section 5, where we compare real demand data with numerical demand data. Finally we perform a sensitivity analysis on the values of input parameters in Section 0 to test the robustness of the system.

Rockwell Arena Software version 12.0 is used for building the simulation model. The simulation model is a representation of the real problem setting. We modeled the inventory management of a single ATM. Real historical demand of ATMs from Dutch commercial banks from January 1, 2008 to April 1, 2009 are used in simulation. The run length is nineteen months of which the first fifteen months are used to warm up and forecast demand. Demand forecasts for the subsequent four months are derived to allow for comparison with real demand. During these months we let the model decide about the order frequency and order quantities. One measurement time interval (See Section 2 for a definition) is assumed. In other words, we keep track of the realized fill rate and costs within the last four months.

In order to simulate different weights for different types of missed demand, we vary between two weights in the simulation; either a high or a low weight. Weights are assigned randomly to days such that every day has either a high or low weight. The assigned weight of potentially missed demand may alter more or less frequently and also a larger or smaller differentiation in weights may be desired in practice. In this simulation study however, the emphasis is on demonstrating that assigning different weights adds value. To avoid complexity we choose to randomly assign weights per each day. We introduce the parameter 'diversity of weights' to indicate the actual values of high and low weights. The 'diversity of weights' is a real value between zero and one. Instead of a standard weight of 1, a high weight is assigned 1 plus the 'diversity of weights' and a low weight is assigned 1 minus the 'diversity of weights'.

Table 2 indicates the values of the input parameters used in the baseline scenario. The value of the 'diversity of weights' is set to 0.8 which means missed demand with high impact weighs 1+0.8=1.8 and missed demand with a low impact weighs 1–0.8=0.2. We argue that this is a reasonable value for the baseline scenario given the fact that ATM users perceive service quality at ATMs differently depending on the respective ATM and the time of ATM usage.

Table 2: Input parameter values for the Baseline scenario

Customer demand = Real or numerical demand	Interest rate = 5%
Reorder moment = 11.00AM	Cost per replenishment = € 120
Standard lead-time = 2 days	ATM capacity = € 260,000
Delivery days = Monday until Saturday	Target fill rate = 98%
Delivery time = 10% between 8-9AM, 10% between 9-10AM, … , 10% between 5-6PM	Degree of risk = 0.5
	Diversity of weights = 0.8

5 RESULTS

This section presents the results of the simulation experiments. We conclude that for all experiments in this paper 200 replications per simulation experiment is sufficient (Law and Kelton 2000).

We animated the inventory and inventory position over time in a line chart for validation purposes. A screenshot of this chart during one replication of the baseline scenario is depicted in Figure 4 to visualize the outcome of the model. As depicted, we started every experiment with an initial inventory level of € 260,000. Figure 4 clearly shows that the inventory level reduces due to customer demand. The horizontal axis depicts the time in number of hours, the vertical axis the inventory in Euros. The red line indicates the inventory and the green line the inventory position over time. Figure 2 demonstrates that the actual lead-time between ordering and delivery varies in length due to the fact that a Sunday is not a working day. Also, the time between deliveries differentiates which is caused by demand uncertainty. The chosen order quantities are neither constant; sometimes the ATM is stocked up to capacity and sometimes the ATM is replenished up to, for example, only € 150,000. A small order quantity might be chosen because it is inefficient to order more. For example, if the weekend demand cannot be covered with a single replenishment, another replenishment is necessary just before the weekend. In this case, the inventory management system decides to put enough cash in the ATM to sustain until just before the weekend.

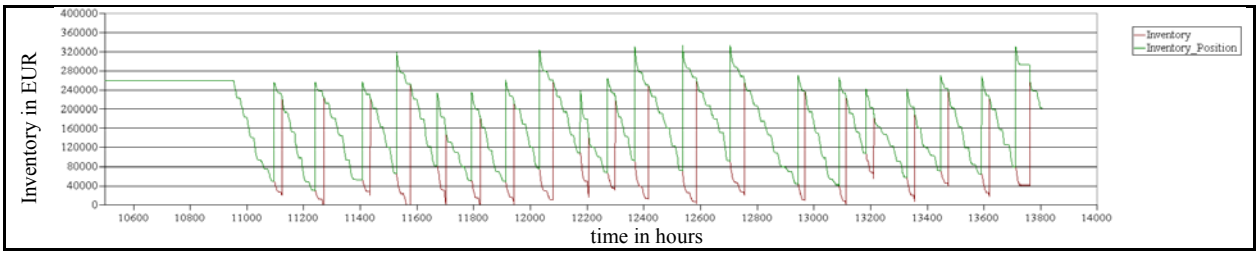

Figure 4: Screenshot of the line chart in the simulation model indicating inventory over time

The main goal of the simulation experiments is to test whether implementing the real-time performance measurement system reduces total cost (i.e., holding and ordering costs) while the perceived service by customers remains unchanged. We performed forty-four experiments in total; eleven similar experiments in four different settings. The four settings are combinations of either real or numerical demand and either with or without the real-time PMS. For each setting we performed eleven experiments indicating various input target fill rates. The results are shown in Figure 5.

The results show that the target fill rate has been achieved for each of the experiments, both for the real, as for the numerical demand. Even for the highest target fill rate (99.5) the experiments show a significant difference with the realized fill rate (99.60, 99.65, 99.71, and 99.68).

All graphs of Figure 5 show that the total cost is exponentially increasing with higher target fill rates. This makes perfect sense because of the stochastic nature of the system (e.g. variability in demand and lead-times) makes it hard and thus expensive to reduce downtime to a minimum.

While differentiating between the weights of missed transactions (i.e., with real-time PMS), we kept track of the regular realized fill rate as well (see Graph b and d in Figure 5; line with diamonds). It turns out that the ATM is unavailable more often when our real-time PMS is integrated (see the regular fill rate in Graph b and d in Figure 5) . However, ATM users are equally satisfied because the adjusted fill rate has been reached (see Graph b and d in Figure 5). The reasoning behind this observation is the following: With the real-time PMS, the majority of missed demand takes place when missed demand has a low weight and less demand is missed when customers assign more weight to ATM unavailability.

In Section 3 we hypothesized that the integration of our real-time PMS either reduces cost or increases the fill rate. If we compare the incurred cost from both settings (i.e., with or without real-time PMS, see Graph c and d in Figure 5), we find proof for Hypothesis 1; an independent two-samples t-test shows a significant cost reduction of 3.7% on average when the real-time PMS is integrated in the inventory man-

agement system. The difference of 3.7% is found when we compare the obtained cost levels in both settings for each input target fill rate and take an average. We find proof for Hypothesis 2 as well; a significant increase in realized fill rate is achieved with equal expenses. For example, when aiming for an adjusted fill rate of 99% (see Graph d), equal costs are incurred as in the scenario of aiming for 98% of regular fill rate (see Graph c).

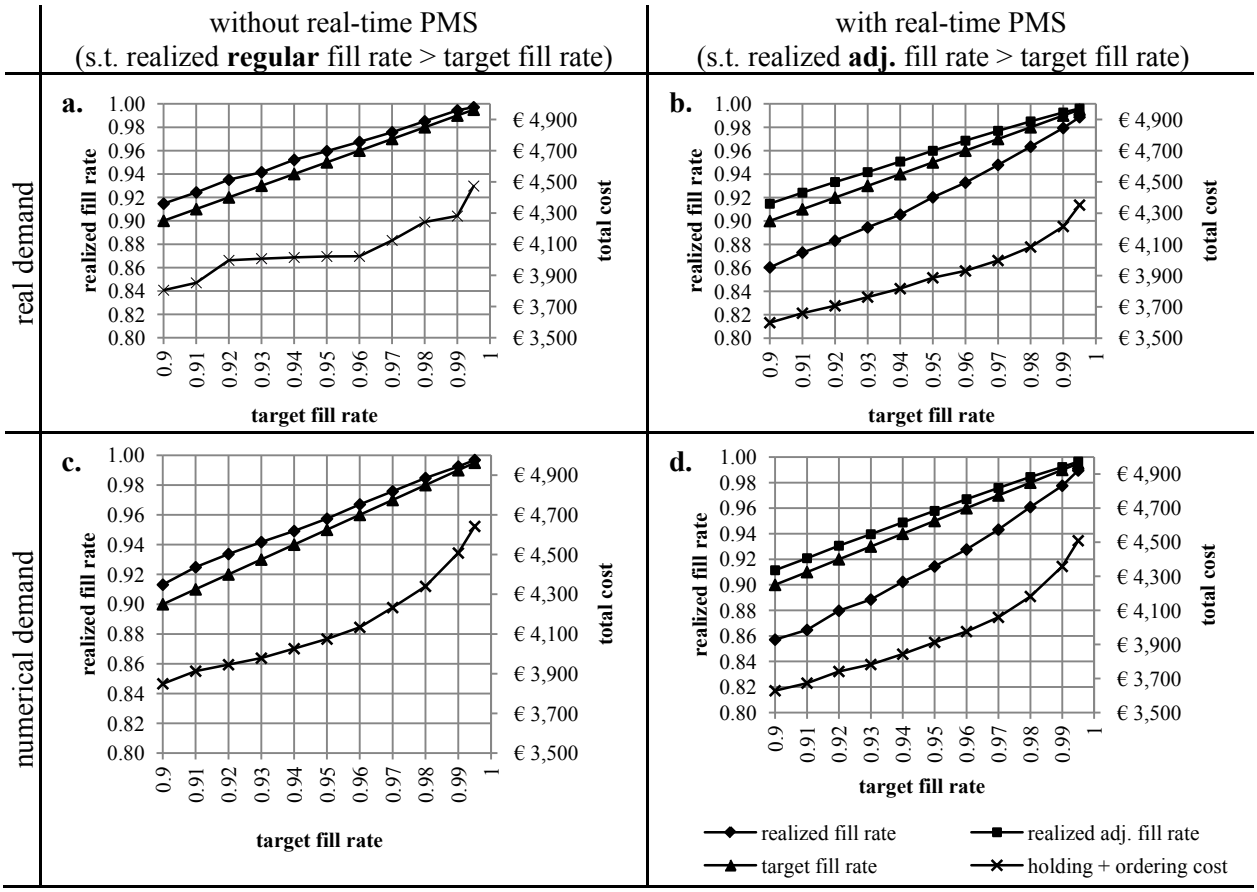

Figure 5. Simulation results depicted in four graphs: a) real demand without real-time PMS, b) real demand with real-time PMS, c) numerical demand without real-time PMS, and d) numerical demand with real-time PMS.Sensitivity Analysis

6 SENSITIVITY ANALYSIS

A genuine simulation model needs to be validated with a sensitivity analysis on input parameters (Kleijnen and Smits 2003). This section provides a sensitivity analysis on the key input parameters (presented in Table 2) by conducting four types of experiments:

a. When considering real demand in Section 5, the demand of a single ATM was considered. Simulation experiments are performed for an additional nine ATMs while considering real demand and baseline parameters. Real demand of the additional nine ATMs is also obtained from Dutch commercial banks.
b. We introduced the input parameter 'degree of risk'. The degree of risk represents the fraction of the 'buffer' (i.e., the demand that can still be missed in order to fulfill the target fill rate) that is allowed to be claimed when deciding about postponing an order. See Section 3 for a detailed description of this parameter. The degree of risk is set to 0.5 in the baseline scenario. To test the ef-

fect of different values on the performance, we performed various experiments while assuming numerical demand.

c. We varied with the cost per order and the interest rate to assess their impact on overall cost.
d. The input parameter 'diversity of weights' is introduced in Section 4. This parameter determines the diversity between the assigned weights to potentially missed demand. We are interested in understanding what happens to the ATM performance with a vast diversity of weights compared to a narrow diversity of weights.

A graph for each type of experiment is shown in Figure 4. From graph a we conclude that the target fill rate of 98% is achieved for almost every ATM while focusing on the regular fill rate and on the adjusted fill rate. The only exception is ATM 8, for which the target fill rate is not achieved. When identifying the cause it seemed that ATM 8 suddenly received much more demand. Apparently the demand forecast did not foresee this sudden increase in demand and this caused too much ATM unavailability in the end. Irregularities and uncertainties in real life make it impossible to ensure the target fill rate is achieved for every individual ATM for every moment in time. Next to that, we strive for a solution that is affordable as well. In reality, ATM cash managers are appointed to achieve a certain fill rate for their entire ATM network and not for each individual ATM. If we average the realized fill rates of all ten ATMs, the realized fill rate is well above the targeted 98%. In practice, this sudden increase in demand could have been noticed earlier and the demand forecast could have been adjusted accordingly. Also, this example illustrates the necessity of a real-time approach.

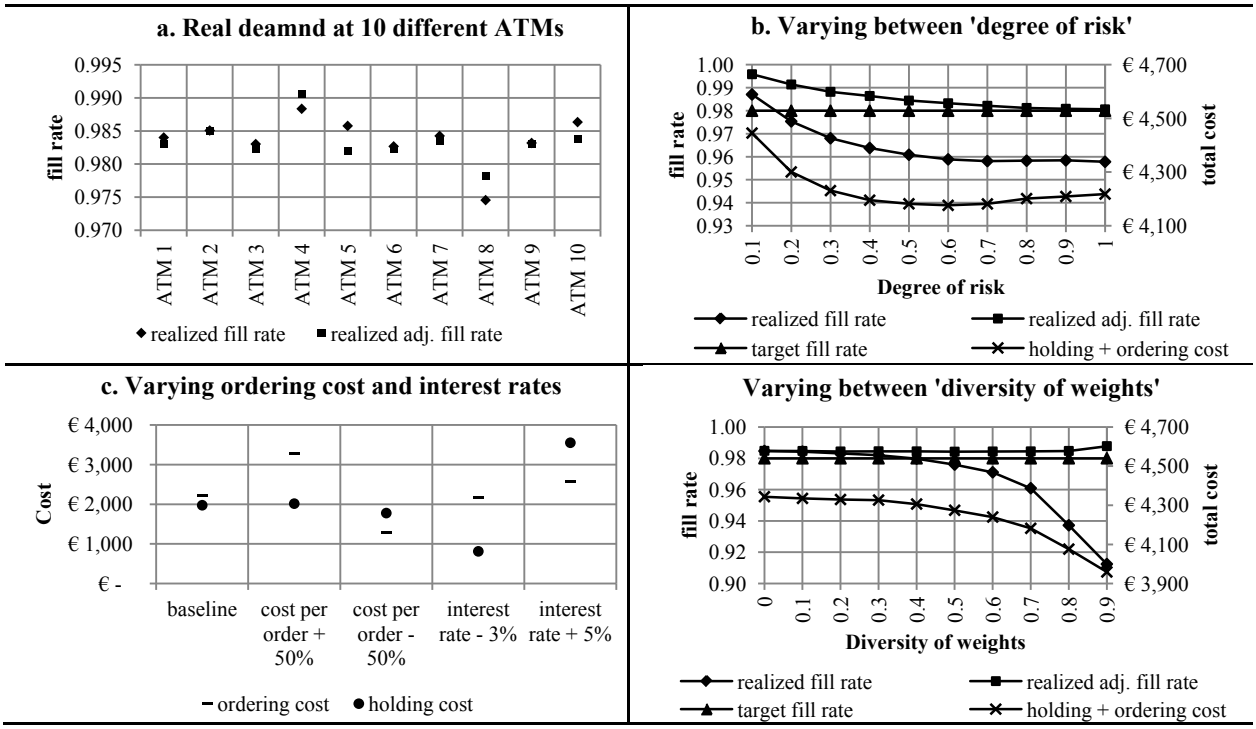

Figure 6: Sensitivity analysis: Four types of experiments

The 'degree of risk' is tested in Graph b in Figure 6. We stated in Section 3 that a low degree of risk is costly because orders would be placed (too) early. Contrarily, a high degree of risk leads to blowing the entire buffer in the beginning of the measurement time interval. When the latter occurs, the system has to account for more safety stock at future orders, which is expensive as well. These expectations are proven valid by the experiments depicted in Graph b. Risk aversion (i.e., a low degree of risk) leads to a higher

fill rate and is indeed expensive (€ 4,445 with a value of 0.1). The higher realized fill rate (99.6% with a value of 0.1) is probably valued by the users, but performing better than agreed does not generate extra income for the manager. When choosing an average degree of risk (degree of risk = 0.5) a total cost of € 4,181 is incurred. The cost increases when choosing a higher degree of risk (€ 4,218 with a value of 0.9), because a large share of the buffer is blown in the beginning of the measurement time interval.

The third graph (see Graph c of Figure 6) checks the sensitivity of two parameters; the cost per order and the interest rate. The first experiment is the baseline scenario and we compare the performance of the other four experiments with the baseline performance. A 50% higher cost per order leads to higher ordering cost and equal holding cost. Someone might expect to see higher holding cost as well, because it makes sense to keep more cash in stock to reduce the order frequency, but in this case the ATM capacity constraints the order quantity. In the third experiment we check the consequence of 50% lower cost per order. As expected, both the ordering cost and the holding cost drop. In the fourth and fifth experiment we vary with the interest rate. An interest rate of 2% instead of 5% reduces the cost of cash and therefore larger order quantities can be expected. However, due to the ATM capacity, the ordering cost remains unchanged. The last experiment in Graph 5 shows that an interest rate of 10% instead of 5% increases both holding and ordering cost. Apparently more orders are being placed with smaller order quantities. All experiments depicted in graph c of Figure 4 represent expected outcomes.

The final set of experiments concerns the input parameter 'diversity of weights' (results are shown in graph d in Figure 6). The total cost is lower when the diversity of weights is larger. For example, a 'diversity of weights' value of 0.9 incurs a total cost of € 4,075 and a diversity of weights of 0 incurs a total cost of € 4,343. As expected, the system allows customer demand to be missed easier when this missed demand does not have much impact on customer satisfaction. Important to notice is that a diversity of weights of 0 equals the scenario of assuming equal weights for missed transactions. Hence, the adjusted and regular realized fill rates are equal as well.

This sensitivity analysis demonstrated that our expectations and assumptions were correct. We also demonstrated the importance of a real-time approach and we showed that the real-time PMS is capable of dealing with a great variety of data.

7 CONCLUSIONS

A real time performance measurement system (PMS) is designed in this paper. We demonstrate that the integration of our real-time PMS into an ATM inventory management system adds value by means of discrete event simulation experiments. The integration leads to a higher fill rate, 99% instead of 98% with equal expenses, or a 3.7% cost reduction while maintaining a 98% fill rate.

The assignment of different weights to missed demand is the key element of our real-time PMS. This property results in a high service quality delivery only when there is also a demand for high quality service. With our real-time PMS in place, ATM users will experience less unavailable ATMs when it really matters, for example, during a national holiday. When ATM unavailability does not have much impact, for example, when many other ATMs are available in the area, ATMs will be out of service more often.

Although we did not integrate the PMS into other processes that are concerned with ATM (un)availability, we believe the real-time PMS can add value in those other areas as well. For instance, integrating our PMS into the process of ATM preventive and corrective maintenance would result in aiming for shorter service windows when the impact of missed demand is higher, and would result in aiming for longer service windows when the impact is lower.

An elaborate sensitivity analysis shows the real-time PMS is robust and is capable of dealing with a great variety of data. In our PMS, the parameter 'diversity of weights' indicates the extent to which the weight of missed demand varies. We show that a larger diversity between weights improves fill-rates or efficiency even further. However in practice, the actual diversity of weights of missed transactions depends on the ATM network. Our measurement system would perform well in especially dynamic environments where the demand for ATM services is highly unstable.

REFERENCES

Cagnazzo, L., P. Taticchi, and A. Brun. "The Role of Performance Measurement Systems to Support Quality Improvement Initiatives at Supply Chain Level." *International Journal of Productivity and Performance Management* 59, no. 2 (2010): 163-85.

Capgemini, RBS, and EFMA. "World Payments Report 2011." 7 (2011): 1-60.

G4S Cash Solutions. "Cash Report 2011." (2011): 1-48.

Ghalayini, A.M., J.S. Noble, and T.J. Crowe. "An Integrated Dynamic Performance Measurement System for Improving Manufacturing Competitiveness." *International Journal of Production Economics* 48, no. 3 (1997): 207-25.

Harvey, Jean. "Service Quality: A Tutorial." *Journal of Operations Management* 16 (1998): 583-97.

Kleijnen, J.P.C., and M.T. Smits. "Performance Metrics in Supply Chain Management." *Journal of the Operational Research Society* 54 (2003): 507-14.

Law, A.M., and W.D. Kelton. *Simulation Modeling and Analysis*. Boston: McGraw-Hill, 2000.

Sargent, Robert G. "Verification and Vaildation of Simulation Models." *Proceedings Winter Simulation Conference, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.* 37 (2005): 130-43.

Silver, E.A., and H.C. Meal. "A Heuristic for Selecting Lot Size Quantities for the Case of a Deterministic Time-Varying Demand Rate and Discrete Opportunities for Replenishment." *Production and Inventory Management Journal* 14 (1973): 64-74.

Suri, P.K., Dilbag Singh, and Ramesh Chander. "A Simulator for Forecasting the Location of Atms in Banking Industry." Paper presented at the Challenges & Opportunities in Information Technology, Punjab, India, 2007.

Van Anholt, R. G., and Iris F. A. Vis. "An Integrative Online Atm Forecasting and Replenishment Model with a Target Fill Rate." *Proceedings of The International Conferene on Logistics and Maritime Systems* (2010): 1-10.

Van Anholt, R.G., and Iris F. A. Vis. "Key Determinants of Atm Performance: Survey Results from Businesses and Users." *Working Paper VU University Amsterdam* (2012).

Wagner, M. "Analyzing Inventory Routing Costs." Paper presented at the International Conference on Enterpreise Systems, Accounting and Logistics, Rhodes Island, Greece, 2010.

AUTHOR BIOGRAPHIES

ROEL G. VAN ANHOLT is a Ph.D. candidate in Logistics at the VU University Amsterdam, The Netherlands. He holds a Bachelor of Engineering degree in Business Engineering from the Professional University Utrecht and a Master of Science Degree (cum laude) in Business Administration (specialization Transport and Supply Chain Management) from the VU University Amsterdam. Next to teaching in Bachelor and Master programmes, he currently develops efficient tools and models to improve cash supply chain performance by applying analytical quantitative research methods in his research. His email address is r.g.van.anholt@vu.nl.

IRIS F.A. VIS is a full professor of Industrial Engineering at the University of Groningen, The Netherlands. She holds an M.Sc. in Mathematics (specialization Operations Research) from the University of Leiden, and a Ph.D. from the Erasmus University Rotterdam. She was an Associate Professor at the VU University Amsterdam and a Visiting Professor at Virginia Polytechnic Institute before joining the University of Groningen. The research interests of Vis are in the design and optimization of container terminals, vehicle routing, supply chain management and inventory management. The common goal in her projects is to develop new planning and control concepts to improve logistics operations by means of techniques from Operations Research.