HARVARD | BUSINESS | SCHOOL

Inventory Management for Mobile Money Agents in the Developing World

Karthik Balasubramanian
David Drake
Douglas Fearing

Working Paper 17-109

Inventory Management for Mobile Money Agents in the Developing World

Karthik Balasubramanian
Harvard Business School

David Drake
Harvard Business School

Douglas Fearing
Los Angeles Dodgers

Working Paper 17-109

Submitted to
manuscript

Inventory Management for Mobile Money Agents in the Developing World

Karthik Balasubramanian and David Drake
Harvard Business School, karthikb@hbs.edu, ddrake@hbs.edu

Douglas Fearing
Los Angeles Dodgers

Mobile money systems, platforms built and managed by mobile network operators to allow money to be stored as digital currency, have burgeoned in the developing world as a mechanism to transfer money electronically. Mobile money agents exchange cash for electronic value and vice versa, forming the backbone of an emerging electronic currency ecosystem that has potential to connect millions of poor and "unbanked" people to the formal financial system. Unfortunately, low service levels due to agent inventory management are a major impediment to the further development of these ecosystems. This paper describes models for the agent's inventory problem, unique in that sales of electronic value (cash) correspond to an equivalent increase in inventory of cash (electronic value). This paper presents a base inventory model and an analytical heuristic that are used to determine optimal stocking levels for cash and electronic value given an agent's historical demand. When tested with a large sample of transaction-level data provided by an East African mobile operator, both the base model and the heuristic improved agent profitability by reducing inventory costs (defined here as the sum of stockout losses and cost of capital associated with holding inventory). The heuristic increased estimated agent profits by 15% relative to profits realized through agents actual decisions, while also offering substantial computational advantages relative to the base model.

1. Introduction

The rapid growth of cellular networks in the developing world in the past decade has laid the groundwork for a potential paradigm shift in financial services for the poor. Traditionally, the poor and unbanked—the roughly one-third of the world's population who do not have an account at a formal financial institution and live on less than \$2 a day—have relied primarily on physical cash when transferring money (Mas 2010). Thus, the velocity of money has been limited by how fast cash can be physically transported, by foot or by bus in most circumstances (Batista and Vicente 2013). This limitation is a critical disadvantage to the poor when money is needed most, such as in the aftermath of a negative economic shock (e.g., sickness or job loss) or a rare opportunity to climb out of poverty through investment (e.g., fertilizer or improved seed purchases) (Helms 2006). At these decisive moments, friends and family willing and able to transfer money have traditionally relied on expensive and/or unreliable cash transfer methods (Morawczynski 2009).

According to the Boston Consulting Group (BCG), the number of unbanked in the developing world with access to mobile phones was estimated to be 2 billion in 2011, and that number is likely to continue to grow (BCG 2011). Recognizing an opportunity, mobile network operators in the developing world began launching money transfer platforms—known colloquially as "mobile money"—with each platform allowing money to be stored and transferred in the form of digital currency (hereafter referred to as e-float). In much the same way that text messages can be sent quickly and cheaply, e-float can be instantly transferred across long distances at a near-zero marginal cost. By the end of 2015, mobile money systems hosted roughly 410 million mobile money accounts in the developing world, a 31% year-over-year increase (Groupe Speciale Mobile Association 2015).

By connecting the poor and unbanked in the developing world to the formal financial system, mobile money has provided several benefits. For example, mobile money has been shown to: enable quicker recovery from economic shocks such as job loss or illness to the primary wage-earner (Jack and Suri 2014); enable more efficient receipt of monetary transfers from non-governmental organizations (NGOs) after disasters (Aker et al. 2011); and lay the foundation for access to formal savings, credit, and insurance opportunities for those who currently lack such access (Mas 2010).

1.1. Transaction mechanics and inventory challenges

Cash-in/cash-out (CICO) agents serve as the backbone of mobile money networks, providing a bridge between physical cash and e-float. These agents, often small shop-owners, invest in inventories of cash and e-float, and then convert cash to e-float ("cash-in" transactions) and e-float back to cash ("cash-out" transactions) for a commission. The following is a typical use case: An urban laborer in Nairobi, Kenya gets paid in cash. He conducts a cash-in transaction with an urban agent in which he gives the agent cash and the agent credits the laborer's mobile money account with e-float (Stage 1 in Figure 1). For her role in executing the transaction, the agent receives a cash-in commission from the mobile money operator (notably, customers generally do not pay for cash-in transactions). The laborer, now with a balance of mobile money, uses his phone to send this e-float to his family outside of Kisumu, Kenya in much the same way he might send an SMS message (Stage 2 in Figure 1). The operator collects a fee from the laborer for executing this person-to-person (P2P) transfer. Having instantaneously received the e-float onto her phone, the laborer's wife goes to the local agent outside of Kisumu to conduct a cash-out transaction. She gives the agent e-float in exchange for cash (Stage 3 in Figure 1). Like the cash-in agent, the cash-out agent is also compensated with a commission from the operator for her role in executing the transaction. Cash-out commissions are generally larger than cash-in commissions, typically by 50% or more. These commissions are generally determined as an increasing step function of transaction value.

Commissions for both cash-in and cash-out transactions are generally paid out from the operator to the agent in e-float monthly (rather than immediately after each transaction). In order to

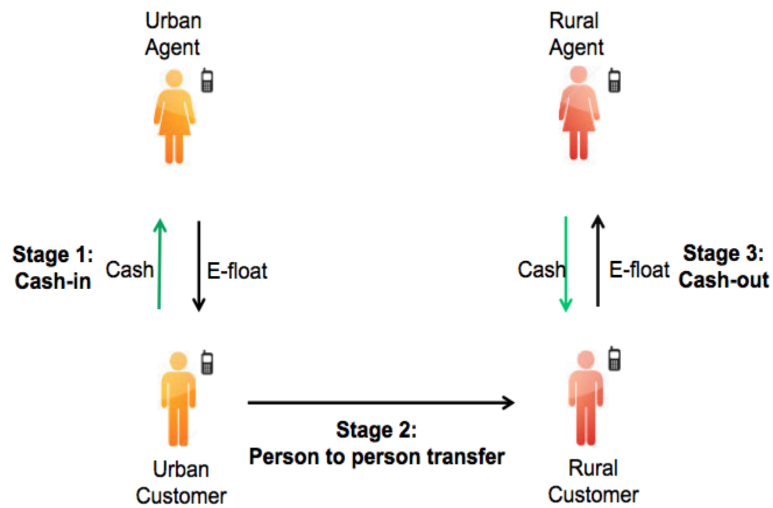

Figure 1 Schematic of a typical urban-to-rural person-to-person transfer, adapted from Agrawal (2009)

conduct a cash-in or cash-out transaction, the agent must have inventory of e-float or cash, respectively. However, stockouts are an acute problem in mobile money networks; agents often run out of cash and e-float. Service levels can fall well below 80% for CICO transactions (Intermedia 2013). Because stockouts of cash and e-float make it harder for customers to easily convert between the two forms of money, they degrade consumer confidence in the convenient convertibility of e-float. Note that because e-float is actual currency, it cannot be "created" on the spot by either the agent or the operator. Each unit of e-float an operator issues must be backed by traditional deposits at a prudentially regulated financial institution. Though moving e-float once it has been issued is clearly easier than moving cash, agents can, and do, nonetheless stock out of e-float.

In managing CICO transactions, the agent's fundamental challenge is an inventory problem: determining how much cash and e-float to carry in order to most-profitably support their mobile money business. This is a non-trivial challenge. In this setting, the agent not only serves uncertain demand for cash and e-float, but each sale of cash (e-float) also generates equivalent inventory of e-float (cash)—i.e., agents not only face stochastic demand, as is typical in many inventory settings, they also face stochastic replenishment for each good through sales of the other good.

1.2. Preview of results

This paper first develops a base model describing the agent's inventory evolution and related costs to determine optimal values of cash and e-float inventory given the agent's stochastic demand. This base approach proves to be too computationally intensive to deploy at scale in the developing world, so a heuristic policy is developed. This heuristic solution bears resemblance to the Newsvendor solution. Employing a large dataset of mobile money agent transactions provided by an East African mobile network operator, §4.2 presents a comparison of the models' performance, showing that both the base model and the heuristic significantly increase agent profitability by reducing inventory costs, defined here as the sum of estimated stockout losses and the cost of capital associated with holding inventory. The heuristic improves estimated aggregate agent profits by 15% relative to actual performance, while also offering a substantial computational advantage relative to the base model. §4.4.2 shows that the heuristic recommendation improvements are robust to changes in the cost of capital parameter. Finally, §4.4.3 shows that agents in the sample who are least "balanced" (that is, their sales are skewed towards either cash-in or cash-out), gain the most from following newsvendor heuristic recommendations.

2. Relation to the Literature

This work relates theoretically to inventory management literature as well as contextually to mobile money literature.

2.1. Inventory Management

The mobile money agent's fundamental inventory challenge is informed by decades of work focused on inventory management under demand uncertainty. While, to the best of our knowledge, there has not been analysis on scenarios where satisfaction of demand for one good generates inventory of another, the setting does relate to the "stochastic cash balance problem," which has been the subject of significant research by the operations management community beginning in the 1960s (for example Girgis (1968), Neave et al. (1970), Chen and Simchi-Levi (2009)). The problem is so-named because a bank (or any general firm) has a challenge in managing its inventory of cash: too much cash results in excessive cost of capital, while too little cash incurs some penalty cost, for example the cost of not meeting a banking reserve ratio requirement. The stochastic cash balance problem is different from the standard stochastic inventory problem because demand can be both positive (withdrawals decrease inventory) or negative (deposits increase inventory). Given this fact, the stochastic cash balance problem has been used to study any product that can be returned, contributing to the development of research on reverse logistics (e.g., Fleischmann et al. 1997). As will be shown, the mobile money agent's problem shares this feature of demand spanning both negative and positive values. However, while stochastic cash balance problems focus on the single trade-off between holding too much and too little cash, the mobile money problem deals with two separate but linked sets of trade-offs: both too much versus too little cash as well as too much versus too little e-float. Furthermore, while the stochastic cash balance literature largely focuses on developing continuous review policies, such as a two sided (s, S) policy (e.g., Porteus and Neave 1972, Hausman and Sanchez-bell 1975) that allow for mid-period inventory adjustments, most mobile money agents do not typically have the opportunity to "re-balance" their inventories mid-day. This allows us to focus on two key values for each agent: the optimal amount of cash and the optimal amount of e-float with which the agent should begin each day.

Because the base developed here requires a level of computation that makes it impractical to deploy broadly, an analytical heuristic is developed. The heuristic bears a strong resemblance to the newsvendor solution, a well-studied framework for the inventory management of short-life cycle goods with stochastic demand (Khouja (1999) and Silver et al. (1998) provide comprehensive reviews). The newsvendor model traces its roots from a 150-year old cash logistics problem: Edgeworth (1888) used newsvendor logic to study the daily cash needs of a bank branch. Since then, the newsvendor model has been adapted to study a wide variety of contexts and extensions – ranging from inventory management in the presence of externalities (Netessine and Zhang 2005) to risk mitigation in networks (Van Mieghem 2007), to the relationship of resource flexibility to a firm's optimal capital structure (Chod and Zhou 2013). However, to the best of our knowledge, the newsvendor model has not been applied in a setting in which sales of one good generate inventory of another as is the case with mobile money. This paper also presents results of newsvendor heuristic evaluations that are consistent with theory on the "pull-to-center" effect (e.g., Schweitzer and Cachon (2000), Bolton and Katok (2008), and Bolton et al. (2012)) in the behavioral operations literature.

2.2. Mobile money

The rapid emergence of mobile money has attracted the interest of fields ranging from economics to sociology to public policy. Jack and Suri (2014) study mobile money's social welfare impacts, finding that households using mobile money were significantly more able to smooth consumption after a negative economic shock (e.g., sickness or job loss) than comparable households not using mobile money. Jack and Suri explain this disparity by demonstrating that mobile money users who experienced shocks received more numerous and larger remittances from farther away than their non-user counterparts. Mobile money had the effect of significantly widening and enhancing informal insurance networks; family and friends were able to send users more money more efficiently in times of crisis. Suri and Jack (2016) also find evidence that access to M-Pesa, the dominant mobile money system in Kenya, has lifted 194,000 households (2% of all Kenyan households) out of poverty since its inception in 2007. Mbiti (2011) demonstrates that mobile money's introduction as a tool for sending money was so disruptive to the markets for remitting money that, in many cases, remittance prices fell by over 50% over a six-year period. Mbiti (2011) also shows that, like the adoption of mobile telephony, early adopters tended to be wealthy, urban, and educated. However, as with mobile telephony, mobile money has progressed down-market and geographically widened its reach very quickly. This is particularly relevant in the developing world where over 80% of adults do not have a bank account (Kendall 2011). Indeed, Mbiti and Weil (2013) notes that the number of mobile money agents in Kenya exceeded 25 times the combined total of bank branches and ATMs in the country. Morawczynski (2009) pursued an ethnographic approach to study mobile money's role in empowering women, finding many of the women interviewed reported that using mobile money to store savings significantly reduced the risk of their husbands appropriating their money, thus increasing their financial autonomy. Balasubramanian and Drake (2015) study how service quality and competition are related to demand – finding that average demand increases with both pricing transparency and agent expertise. That study also finds that agent expertise interacts positively with competitive intensity, suggesting that expertise is a significant dimension of competition between agents. Finally, Aker et al. (2011) study Concern Worldwide's (CW) response to the 2010 drought crisis in Niger. Instead of distributing physical relief items, CW distributed money. Each month for 5 months, some beneficiaries received physical cash transfers, and others received money transfers via mobile money. Aker et al. (2011) show that the cost of distributing monetary assistance via mobile money, as well as the cost to the beneficiaries of receiving mobile money was significantly lower than costs associated with physical cash distribution. As more governments and NGOs shift emergency relief from the distribution of goods to the distribution of money, mobile money's importance post-disaster is expected to increase (BTCA 2014). Accordingly, improving mobile money agents' service reliability will also become increasingly important. To the best of our knowledge, this paper is the first to address this challenge. We do so by applying an operations research lens to this context, with an eye toward improving mobile money agents' inventory management.

3. Inventory Models

Mobile money agents face stochastic demand and interrelated stochastic replenishment for two goods—i.e., sales of cash (cash-out transactions) generate inventory of e-float, and sales of e-float (cash-in transactions) generate inventory of cash. As a consequence, traditional inventory models cannot be applied to this context in a straightforward manner. Accordingly, we model the mobile money agent's problem, beginning by developing a base model to derive agents' optimal starting cash and e-float quantities. However, generating daily recommendations for thousands of agents still requires an impractical level of computation. To address this limitation, we develop an analytical heuristic. This section will begin with a description of the parameters common to both models, and then address each model in turn.

3.1. Description of parameters

At the start of each day, an agent facing daily per-unit cost of capital γ chooses a budget b, and then splits this budget between her daily starting cash quantity q_1 and her starting e-float quantity $f_1 = b - q_1$. The agent then experiences a series of N demand arrivals (where "N" is not known ex-ante), represented as a sequence of cash and e-float demand arrivals, D_t^c and D_t^e respectively, where $t \in \{1, 2, ..., N\}$. For all t, D_t^c and D_t^e are non-negative, and cannot both be positive (i.e. each arrival is characterized by either cash or e-float demand). Define p as the probability that any given arrival will be cash demand, and $1-p$ as the probability that any given arrival will be e-float demand. The inventories of cash and e-float (represented as q_t and f_t respectively) thus change throughout the day as demand arrives. When the agent is presented with cash demand in a given arrival t, the agent earns a per-unit commission m_c for all cash sales and receives e-float equivalent in value to those sales. The amount of e-float available in the following period is increased (i.e. $f_{t+1} = f_t + \min(q_t, D_t^c)$. Similarly, when the agent is presented with e-float demand in a given arrival t, she receives per-unit commission m_e for all e-float sales and receives cash equivalent in value to those e-float sales. Thus, the amount of cash inventory in the next period is increased (i.e. $q_{t+1} = q_t + \min(f_t, D_t^e)$. Table 2 lists these parameters and summarizes their respective descriptions. In both of the modeling approaches described in the section, it is assumed (i) that demand realizations are i.i.d., (ii) that the agent does not re-balance during the day, and (iii) that any demand not satisfied is lost to the agent.

3.2. The role of arrival sequencing

Uncertain sequencing of cash and e-float arrivals complicates decision-making in this setting. To illustrate this challenge, take the following example where initial inventories of cash and e-float are $q_1 = 100$ and $f_1 = 100$. If the agent experiences a demand sequence $\{D_1^c = 100,\ D_2^e = 200\}$, all of this demand is satisfied. This is because the e-float quantity increases after satisfying demand for units of cash, resulting in sufficient additional units of e-float available for the next arrival ($f_2 = f_1 + \min(q_1, D_1^c) = 100 + 100$). Now consider the same arrivals, but in the reverse sequence $\{D_1^e = 200,\ D_2^c = 100\}$: only 200 of the 300 units of total demand are satisfied. This is because the agent is only able to satisfy 100 of the initial 200 units of e-float demand. It is possible that the agent can satisfy all demand for cash and e-float with relatively little inventory (potentially far less than the sum of all demand) if the sequencing of that demand is favorable. Figure 2 presents a simplified depiction of inventory positions of cash and e-float over the course of a day for an agent who begins with 100 units of both cash and e-float. In this case, the agent stocks out of both cash and e-float over the course of the same day as a result of unfortunate demand sequencing.

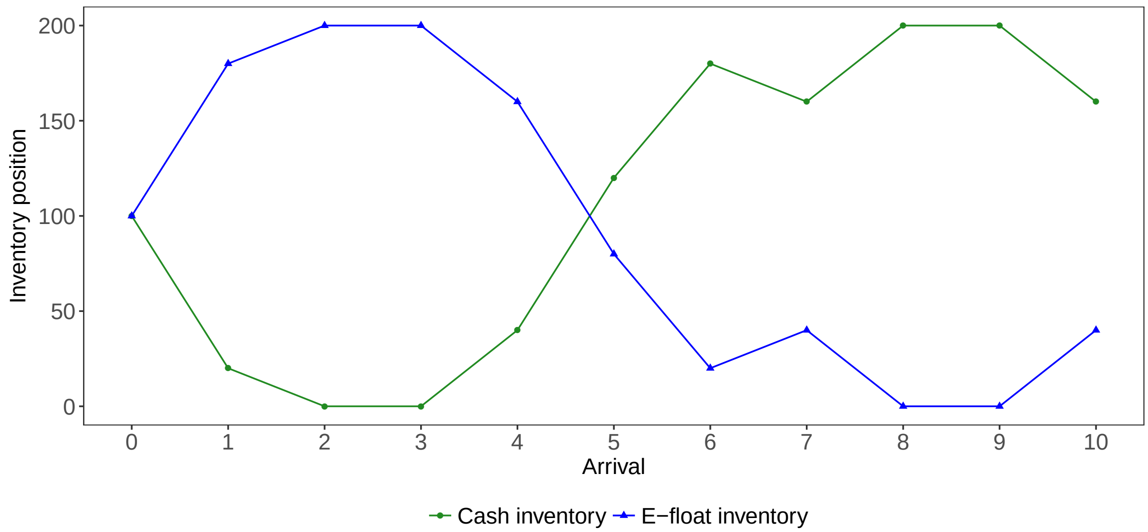

Figure 2 An illustrative example of agent inventory balances of cash and e-float

3.3. Combining cash and e-float demand distributions

The agent's problem can be simplified significantly by combining cash and e-float demand into a single demand variable.

Lemma 1 *Without loss of generality, a single random variable can represent both cash demand and e-float demand. Specifically:* $D_t = D_t^c - D_t^e$.

All proofs, including that for Lemma 1, are provided in Appendix B. The steps that follow summarize the construction of the probability mass function (pmf) of the combined demand distribution from the pmf of cash demand magnitude $f_{D_t^c}(x)$, the pmf of e-float demand magnitude $f_{D_t^e}(x)$, as well as the probability that any given arrival is a cash arrival, p. First, reflect $f_{D_t^e(x)}$ about the y-axis, scale this new pmf by $1-p$, then finally combine with the scaled (by p) pmf of cash demand magnitude. In mathematical form:

$$f_{D_t(x)} = (1-p)\cdot f_{D_t^e}(-x) + p\cdot f_{D_t^c}(x)$$

Through this transformation, e-float demand is represented by negative values of D_t and cash demand is represented by positive values of D_t. Now, the evolution of both cash and e-float inventories, as well as cash and e-float underages, can be characterized with D_t.

3.4. Base model

To formalize and provide the underlying structure of each mobile money agent's inventory problem, a base model which casts the agent's inventory process as a Markov chain is presented here. Through recursion, this base model can be used to solve for the profit-maximizing daily starting values of cash and e-float. While the base model is useful as a functional benchmark, as well as an illumination of the agent's inventory problem, this base approach is not well-suited as a solution for this setting because it is computationally infeasible at scale. With this caveat, we now present the formulation and solution approach of the base model.

The agent's objective is to maximize profits by choosing her optimal budget b^* and the optimal starting cash quantity q_1^*. The optimal starting e-float quantity again follows directly as $b^* - q_1^*$. To identify the optimal beginning budget and inventory positions, the optimal cash quantity for a single budget is found, and then this process is iterated over the range of possible budgets. Given that the number of demand arrivals is uncertain, the state variables need to not only track the current cash position (and thus, implicitly, e-float position) but also an indicator to capture whether the final demand had arrived. This indicator variable is denoted as s_t, taking the value of 1 if the final arrival had not occurred prior to t and 0 otherwise (i.e. if $s_t = 0$, there would be no more arrivals for the day).

3.4.1. Stochastics As described by Lemma 1, let D_t be a random variable characterized by the distribution of the magnitude of demand arrivals (positive for cash demand and negative for e-float demand) for a given agent-day. Assume also that the number of arrivals is a geometric random variable—allowing θ_t to be a binary random variable that takes a value of 1 if the final arrival occurs at t and 0 otherwise. Using these definitions, the state transitions according to the following transition function:

$$f\left(q_t, s_t, b\right) = \langle q_{t+1}, s_{t+1}\rangle = \langle (1-s_t)\cdot q_t + s_t \cdot \min\left(b, \max\left(q_t - D_t, 0\right)\right), s_t \cdot (1-\theta_t)\rangle$$

We also define the geometric success parameter λ, such that $P(\theta_t = 1) = \lambda$. In other words, λ represents the probability that any given arrival is the final arrival. Thus, transition probabilities can be written below:

$$P\Big(\langle q_{t+1}, s_{t+1}\rangle \mid \langle q_t, 1\rangle\Big) = \begin{cases} (1-\lambda)\cdot P(D_t = q_t - q_{t+1}) & \text{if } s_{t+1} = 1, 0 < q_{t+1} < b \\ (1-\lambda)\cdot P(D_t \geq q_t) & \text{if } s_{t+1} = 1, q_{t+1} = 0 \\ (1-\lambda)\cdot P(D_t \leq q_t - b) & \text{if } s_{t+1} = 1, q_{t+1} = b \\ \lambda \cdot P(D_t = q_t - q_{t+1}) & \text{if } s_{t+1} = 0, 0 < q_{t+1} < b \\ \lambda \cdot P(D_t \geq q_t) & \text{if } s_{t+1} = 0, q_{t+1} = 0 \\ \lambda \cdot P(D_t \leq q_t - b) & \text{if } s_{t+1} = 0, q_{t+1} = b \end{cases} \quad (1)$$

$$P\Big(\langle q_{t+1}, s_{t+1}\rangle \mid \langle q_t, 0\rangle\Big) = \begin{cases} 1 & \text{if } s_{t+1} = 0, q_{t+1} = q_t \\ 0 & \text{otherwise} \end{cases}$$

3.4.2. Costs and optimization Letting the maximum number of arrivals be some arbitrarily large integer N, set the terminal reward to $R_N(q_N, s_N, b) = 0$ and define the stage reward function to be:

$$R_t(q_t, s_t, b) = s_t \cdot \left(\mathbb{E}_{D_t}\left[\left(m_c \cdot (D_t - q_t)^+ + m_e \cdot (q_t - b - D_t)^+\right)\right]\right).$$

A recursion can be used to determine the optimal starting quantities of cash and e-float (no decisions are made after the initial quantity selections). The cost-to-go function follows:

$$\begin{aligned} J_t(q_t, s_t, b) &= R_t(q_t, s_t, b) + \mathbb{E}\left[J_{t+1}\left(f\left(q_t, s_t, b\right)\right)\right] \\ \mathbb{E}[J_{t+1}(f(q_t, 0, b))] &= J_{t+1}(q_t, 0, b) = 0 \\ \mathbb{E}[J_{t+1}(f(q_t, 1, b))] &= \sum_{q_{t+1}=0}^{b} \Big(P(\langle q_{t+1}, 1\rangle \mid \langle q_t, 1\rangle) \cdot J_{t+1}(q_{t+1}, 1, b)\Big) \end{aligned}$$

Thus, given the daily cost of capital parameter, γ, the optimal cost for a each given budget, optimal budget and optimal cash share of the optimal budget, respectively, are:

$$\begin{aligned} J_1^*(b) &= \min_{q_1} J_1(q_1, 1, b) \\ b^* &= \underset{b}{\operatorname{argmin}} \ \left(J_1^*(b) + \gamma \cdot b\right) \\ q_1^* &= \underset{q_1}{\operatorname{argmin}} \ J_1(q_1, 1, b^*) \end{aligned}$$

3.4.3. Limitations There are significant practical limitations to this base model. Due to an extremely large solution space per agent-day, the computation time required for each agent-day is prohibitively long. Even while using highly-parallelized computing resources, the computation of the optimal starting quantities of cash and e-float for each agent day takes approximately 164 seconds on average. Given the fact that operators have thousands of agents in their networks, it is unlikely that mobile money platform operators would be able to utilize this base model daily (and agents in this context typically do not have computational resources of their own). Second, the base model requires inputs of both per-agent per-day estimates of discrete demand distributions (the magnitude of each demand arrival) and per-agent per-day estimates of the number of demand arrivals. These estimations may compound, causing a degradation of model performance. Finally, the fact that the base model is a "black box" may limit adoption and usage by mobile money agents because there is no simple, intuitive explanation for how the recommendations were developed.

3.5. Newsvendor heuristic

Due to the base model's practical limitations, we develop a heuristic based on a lower bound on the base model's underage that yields useful results without significant computation. The objective remains to find the agent's optimal daily starting inventory of cash and e-float, (q_1^* and $b^* - q_1^*$, respectively) as a function of demand, the daily cost of capital (γ), and cash-out and cash-in commissions (m_c and m_e).

Given that the full satisfaction of a cash arrival would yield a corresponding increase in e-float inventory, intuition suggests that cumulative demand (or equivalently, net demand) after t arrivals would be an important quantity in this setting. Define $\Delta_t = \sum_{i=1}^{t} D_i$. Now define two additional quantities based on cumulative demand: maximum cumulative demand ($\hat{\Delta}_t = \max_{1\leq j\leq t} \Delta_j$) and minimum cumulative demand ($\check{\Delta}_t = \min_{1\leq j\leq t} \Delta_j$). As in the base model, let q_1, represent the starting cash quantity, and thus let $b - q_1$ represent starting e-float quantity.

Proposition 1.A *If the initial cash quantity is greater than the maximum cumulative demand ($q_1 \geq \hat{\Delta}_t$) and the initial e-float quantity is greater than the negative of the minimum cumulative demand ($b - q_1 \geq -\check{\Delta}_t$) then all demand up to and including arrival t will be satisfied.*

Proposition 1.B *Given that there are both cash and e-float arrivals (i.e. $\exists$ x and y such that $D_x > 0$ and $D_y < 0$), all demand can be satisfied with strictly less cash and e-float inventory than the sum of all cash and e-float demand.*

As will be demonstrated in this section, Propositions 1.A and 1.B are foundational to the development of the newsvendor heuristic described here. The implication of Proposition 1.B is significant: agents can satisfy all demand while stocking less inventory (sometimes substantially less) than the sum of all demand. The intuition for this result is as follows: the arrival of cash demand (and satisfaction of that demand) generates e-float inventory which can be used to satisfy future e-float demand and vice versa. As a consequence, satisfying e-float (cash) demand in one period contributes to satisfying cash (e-float) demand in a future period.

An illustrative example is presented in Table 1, which shows a sequence of cash-in and cash-out arrivals, inventory positions (with initial inventories $q_1 = f_1 = 100$), lost sales, and cumulative demand. In this case, if an agent begins the day with at least the maximum cumulative demand in cash (i.e. starts with at least 120 units of cash) and at least the negative of the minimum cumulative demand (i.e. 140 units or more of e-float), then the agent would be able to satisfy all cash and e-float demand. Thus the agent in this case could have satisfied all 460 units of demand by holding only 260 units of inventory.

While the maximum and minimum cumulative demands are clearly not known ex-ante, they can be represented as random variables. These maximum and minimum cumulative demand distributions are used as the basis for a heuristic inventory policy. Though the number of arrivals an agent will experience in any given day is uncertain, the maximum and minimum of cumulative demand can still be represented for an arbitrary N as $\hat{\Delta}$ and $\check{\Delta}$, respectively (allowing us to drop the "N" subscript).

Arrival, t	1	2	3	4	5	6	7	8	9	10
Demand, D_t	80	30	10	-40	-80	-60	20	-60	-40	40
Cash inventory, q^t	100	20	0	0	40	120	180	160	200	200
E-float inventory, f^t	100	180	200	200	160	80	20	40	0	0
Cash underage, $(D_t - q_t)^+$	0	10	10	0	0	0	0	0	0	0
E-float underage, $(-D_t - f_t)^+$	0	0	0	0	0	0	0	20	40	0
Cumulative demand Δ_t	80	110	**120**	80	0	-60	-40	-100	**-140**	-100

Table 1 Illustrative example of agent demand process and inventory evolution

Proposition 2.A *The positive part of the difference between maximum cumulative demand and the initial cash quantity $(\hat{\Delta}_t - q_1)^+$ is a lower bound on the cash underage of the base model. The positive part of the difference between the negative of the minimum cumulative demand and initial e-float quantity $(-\check{\Delta}_t - b + q_1)^+$ is a lower bound on e-float underage in the base model.*

Proposition 2.B *This lower bound is sharp in all cases except those where the agent experiences both e-float and cash stockouts (when $\exists$ both x and y such that $D_x > q_x$ and $-D_y > b - q_y$).*

Thus, for the purposes of developing the heuristic, this lower bound on underage is treated as the underage itself. The overage will be captured by incorporating a daily capital cost, γ, which penalizes inventory holding. The cost of capital in this setting is analogous to the cost to acquire (build or purchase) inventory in most other newsvendor settings. A unit of cash or e-float costs the agent γ in capital cost (interest or the lost opportunity to deploy capital toward other ends). This capital cost is not salvageable. The cost function, then, can be written as:

$$G(q_1, b) = \mathbb{E}\left[m_c \cdot \left(\hat{\Delta} - q_1\right)^+\right] + \mathbb{E}\left[m_e \cdot \left(-\check{\Delta} - (b - q_1)\right)^+\right] + \gamma \cdot b \tag{2}$$

This cost function is convex in b and q_1 (the proof is provided in Appendix B). Therefore, first order conditions can be generated to determine the cost-minimizing q_1 and b as a function of the cost of capital γ, cash commission m_c, e-float commission m_e, the distribution of maximum cumulative demand, $F_{\hat{\Delta}}(\cdot)$, and the distribution of minimum cumulative demand, $F_{\check{\Delta}}(\cdot)$.

Proposition 3 *An agent's optimal starting values of cash and e-float for the newsvendor heuristic are: $q_1^* = \left(F_{\hat{\Delta}}^{-1}\left(1 - \frac{\gamma}{m_c}\right)\right)^+$ and $b^* - q_1^* = f_1^* = \left(-F_{\check{\Delta}}^{-1}\left(\frac{\gamma}{m_e}\right)\right)^+$ respectively.*

As intuition would suggest, the heuristic recommends holding less cash and e-float as the cost of capital increases, and more cash and e-float when the commission for cash and e-float sales, respectively, are greater.

While the newsvendor heuristic is developed from a lower bound on underage cost, another heuristic can be developed from an upper bound on underage cost. Appendix A.1 details how framing the agent's problem as a stochastic flow problem approximates the agent's inventory evolution as a reflected Brownian motion (RBM). This allows for the development of another heuristic that provides an upper bound on the base model's underage cost (see Proposition 4 in Appendix A). This RBM heuristic also performs well relative to actual agent's decisions but, as shown in Figure A.1, it under-performs relative to the base model and the newsvendor heuristic.

Parameter	Model	Description
γ	Both	daily unit cost of capital
m_c	Both	per-unit commission on cash sales
m_e	Both	per-unit commission on e-float sales
t	Both	arrival number
q_t	Both	cash inventory at arrival t
f_t	Both	e-float inventory at arrival t
q_1	Both	beginning of day cash inventory
f_1	Both	beginning of day e-float inventory
D_t	Both	value of demand at arrival number t
s_t	Base	1 if final arrival has not yet occurred by arrival t
θ_t	Base	1 if arrival t is final arrival
λ	Base	probability of any given arrival being the final arrival
Δ_t	Heuristic	cumulative demand after t arrivals
$\check{\Delta}_t$	Heuristic	minimum cumulative demand after t arrivals
$\hat{\Delta}_t$	Heuristic	maximum cumulative demand after t arrivals

Table 2 Table of model parameters

4. Performance Evaluation with Historical Data

Using historical data as input to each of the models, recommendations of starting values of cash and e-float inventories per agent per day can be generated, and the models' performance can be compared to agents' actual performance. This section conducts such an analysis, beginning with a description of historical data, followed by a description of the results from the evaluation of the base model and newsvendor heuristic against actual decisions for all agents over the final 6 months of the sample period. Additional findings of interest related to the newsvendor heuristic are presented: a comparison of the models at an agent-level, a sensitivity analysis with respect to the cost of capital parameter, and an analysis of how "balance" (share of cash-in versus cash-out sales) of an agent affects the performance differential between models.

4.1. Historical data

To evaluate agents' actual decisions against recommendations, a large sample of transaction-level cash-in and cash-out historical data from a scaled mobile money operation in East Africa was utilized. Real Impact Analytics, a company specializing in business analytics for mobile networks in the developing world, provided anonymized transaction logs, each of which features a time-stamp, transaction type, transaction value, anonymized IDs of the sender and receiver, as well as the pre and post e-float balances of both the sender and receiver. These transaction logs contained all 35,882,460 transactions from 6,725 agents who conducted at least one cash-in and one cash-out transaction on at least 336 of the 471 days (roughly a 5 day work-week over the period) between June 1, 2014 and September 14, 2015. The total number of agent-days in the sample was 2,708,385. With this data, each agent's actual daily e-float decision as well as all e-float (cash-in) and cash (cash-out) sales can be directly identified. Because neither cash inventory nor stockouts are directly observable in the data, these quantities were estimated as described below.

4.1.1. Estimating cash inventory While historical e-float balances (and thus, agent e-float stocking decisions) are directly observable over time in the transaction data, cash inventory is not directly observable in the data. In order to estimate historical cash decisions made by agents, agent-level e-float balances over time are utilized in two-week blocks. Two assumptions are made. First, that each agent's budget does not fluctuate over a two-week block; and two, each agent has stocked out of cash (and thus has allocated all of her budget towards e-float) at some point over the two weeks. Under these two assumptions, the maximum value of e-float balance represents the agent's inventory budget. From the budget estimate, the agent's cash balance at any given time can be estimated by subtracting the current e-float balance from the maximum balance over the two week period prior to any given day. Because the assumption that an agent's budget does not fluctuate over time is not likely to hold, it is helpful to observe that, if that assumption is violated, our budget estimates are over-estimates. Because this process likely over-estimates the agent's cash inventory at any given time, only cash-in revenue and cost of capital attributable to e-float holding are compared across the models; this is discussed further in the discussion section. Our results showing that the newsvendor heuristic performs significantly better than actual agent decisions are robust to the choice of the duration used to calculate and agents budget; using a week, a month, and the entire sample period as this duration (rather than two weeks) all yield similar results.

4.1.2. Reconstructing censored demand Due to demand censoring, historical sales data are not equivalent to historical demand. There are various methodologies proposed in the literature to derive demand from sales (van Ryzin and Talluri (2005) provides an description of many common methodologies). However, these methods, in general, focus on scenarios where firms are unable to satisfy demand until the next period once a stockout has occurred. These approaches are not applicable in the mobile money setting. In the mobile money setting, inventory of e-float (cash) can be generated mid-period from sales of cash (e-float) — making it possible for stockouts to occur and be resolved within a given day. For this reason, we opt to employ a simple three-step imputation process to estimate demand. To recreate total historical demand for each agent-day, we 1) estimate the timing and duration of stockouts of e-float (cash), 2) estimate e-float (cash) demand arrival rates, and 3) estimate the magnitude of e-float (cash) lost sales to insert in each e-float (cash) stockout interval.

In the first step, we estimate the timing and duration of stockouts. We define the e-float (cash) stocked out state as any interval within which an agent is holding less than a threshold amount of e-float (cash) inventory. This threshold is agent-specific and was chosen as half of each agent's mean e-float (cash) transaction size over the sample period.[1] Because transactions are time-stamped with pre and post balances, the time and duration of each estimated e-float stockout interval can be exactly determined. Similarly, with the assumption of fixed budget allowing us to infer cash balances at all points in time, the time and duration of each estimated cash stockout interval can also be determined. In the second step, we estimate an arrival rate for each agent-day. To generate these arrival rates, we calculate the average arrival rate (of units of cash and e-float demand, respectively) for each agent for each day of the week, using only days that the agent had inventory levels of cash and e-float above the stockout thresholds for the entire day. This is done because the day of the week is a significant factor in demand arrival rate. In the third step, we estimate the magnitude of demand arrivals to insert into each stockout interval. For each duration of stockout, we estimated this quantity by taking the product of the duration of the stockout and the arrival rate of e-float (cash) demand.

4.2. Model-specific inputs and parameter estimation

In this subsection, the estimation of parameters specific to each model is described. The cost of capital parameter is common to both models; the value of this parameter and a sensitivity analysis is presented in §4.4.2.

4.2.1. Base model The base model requires estimates of both the geometric success parameter λ (to account for uncertainty in the number of arrivals) as well as a transition matrix to describe the probabilities of transitioning between states. First, λ is calculated from an estimate of the number of arrivals for each agent-day. This estimate of the number of arrivals is generated with a one-step seasonal point forecast using the number of arrivals seen by that agent on all previous days. The seasonal forecasting method accounts for day-of-week effects on arrivals. Second, the transition matrix (as specified in equation 1) is generated from an empirical demand distribution of the magnitudes of each agent's historical sales.

[1] The results presented in this paper are robust to changes in threshold; using the mean, a fourth of the mean, and median do not produce meaningfully different results.

4.2.2. Newsvendor heuristic The newsvendor heuristic requires only estimates of the distributions of maximum and minimum cumulative demand. One-step seasonal forecasts of minimum and maximum cumulative demand to generate means and standard deviations of normal distributions, which are used as the forecasted maximum and minimum cumulative demand distributions, respectively. The newsvendor heuristic fractiles are applied to these respective distributions to generate each agent's optimal budget and their starting cash and e-float inventory.

4.3. Comparison of model results

To evaluate the performance of each of the models, the transaction logs for each agent each day are used, augmented with re-created demand arrivals generated using the process described above.

We evaluate model and actual performance over the final 180 days of transaction data, generating recommendations from the base model and newsvendor heuristic for each agent-day within this horizon. Each model generates recommended starting cash and e-float inventory for each agent-day. For each set of values (base model and newsvendor recommendations, as well as the actual e-float decision and estimated actual cash decision) we simulate the day using sequenced demand from the augmented transaction log. The commission lost from unmet demand is summed to generate a stockout loss estimate for each model for each agent-day. The cost of capital associated with e-float inventory holding each agent-day is also calculated. As stated earlier, only the stockout losses from cash-in (e-float demand) transactions and cost of capital associated with e-float are presented here, as the agent's actual cash decisions are not precisely calculable. These results are aggregated across all agents over the 180-day period. A performance comparison of agents' estimated actual decisions, newsvendor heuristic, base model, and hindsight decisions are presented in Figure 3. The hindsight results are calculated by determining the cost of capital associated with stocking the minimum inventory required to satisfy all sequenced demand. Thus the hindsight decisions result in no stockout losses, but some cost of capital.

Each bar in Figure 3 represents the total amount of commission that could have been earned had all agents satisfied all e-float demand. The red portion represents e-float commission lost due to e-float stockouts, the orange represents the cost of capital allocated to e-float, and green represents the net revenue realized by all agents in each model. As can be seen, there are striking differences in performance across the actual, heuristic, base model, and hindsight decisions. Paired t-tests reveal statistically significant differences in total e-float inventory costs (stockout + cost of capital) between the NV heuristic and actual decisions, as well as between the base model and actual decisions. Both of these tests result in values of $p < .001$. The newsvendor heuristic results in e-float stockout losses as a percent of total possible e-float commission that are nearly 15 percentage points (20.8% to 5.9%) less than those that correspond with the actual decisions made by agents,

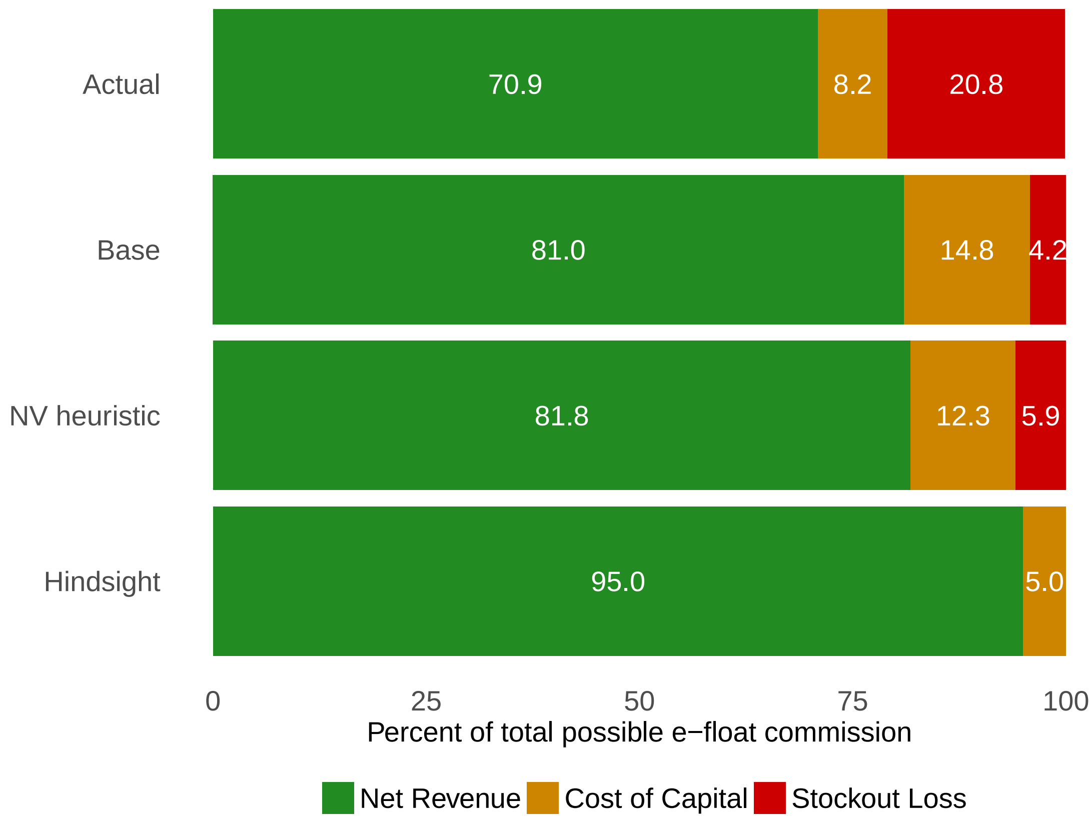

Figure 3 Performance evaluation: aggregate net e-float revenue for actual agent decisions, base model, newsvendor heuristic, and hindsight recommendations over 180 days

while only requiring 4.1 percentage points of extra capital cost (8.2% to 12.3%). The resulting 10.9 percentage point increase of the newsvendor heuristic net revenue over actual net revenue is a key result of this paper. While the net revenue in the hindsight optimal scenario is 13.2 percentage points, greater than the newsvendor heuristic, the exact sequencing and magnitudes of demand arrivals must be known ex-ante to achieve this level of improvement. Some of this performance gap between the newsvendor heuristic and hindsight optimal decisions can likely be closed through more accurate forecasts for the minimum and maximum cumulative demand distributions (upon which the newsvendor fractiles are applied), which is left for future work. Results presented in Figure 3 assume an annualized cost of capital (γ) of 20%; an analysis of how the heuristic performs relative to actual decisions as a function of cost of capital is presented in §4.4.2.

Note that the newsvendor heuristic slightly out-performs the base model. This is possibly due to the base model's requirement of two sets of separate inputs (estimates for both the number of arrivals and the distributions of demand magnitude per arrival); the compounding of estimation errors may be hampering performance. The newsvendor heuristic, on the other hand, requires only a single set of inputs (the distributions of the maximum and minimum cumulative demand). It is also noteworthy that the heuristic's stockout losses are higher than the base model (5.9% to 4.2%), while the capital costs are lower (12.3% to 14.8%). This observation is a direct result of the fact that the newsvendor heuristic is built upon the lower bound on underage cost – which leads to lower inventory recommendations than the base model. This lower inventory recommendation results in both higher stockout losses and lower capital costs. Another heuristic using reflected Brownian moution (RBM) is described in §A.1. Because this RBM heuristic is developed from an upper bound on underage cost, the RBM model yields, as expected, lower stockout losses but higher working capital costs than the newsvendor heuristic and base model, as presented in Figure A.1. However, the RBM heuristic under-performs both the base model and the newsvendor heuristic.

The results presented are robust to changes in the 180-day evaluation horizon: comparisons using an evaluation horizon of the final day, week, month, and three months of transaction data do not yield materially different results.

4.4. Additional NV heuristic analyses

Further analysis is focused on the newsvendor heuristic. Here we show explore performance at an agent-level, as well the heuristic's performance as the cost of capital parameter γ is varied.

4.4.1. Agent level performance At an agent level (as opposed to aggregate savings presented in Figure 3), most agents see significant benefit from the newsvendor heuristic. As seen in Figure 4, while a small subset (roughly 10%) would be worse off (none by more than 20 percentage points), the vast majority of agents could increase net revenue by using the heuristic's recommendations. The mean agent increase is 9.9 percentage points, and the median agent increase is 8.2 percentage points. This discrepancy arises because some agents would benefit substantially more, with the largest increase in net revenue for an individual agent at more than 50 percentage points.

4.4.2. NV heuristic sensitivity to cost of capital Based on informal interviews with mobile money agents, the daily cost of capital parameter γ was estimated and assigned a value of 0.05% (20% annualized). Agents were asked about the terms of the credit they had received (if they borrowed money to finance their inventory) or the terms of credit they would be willing to extend (if they had financed inventory without borrowing). While there was a range of responses between 5% and 80% annualized, 20% seemed to be the most representative single number. However, given that the cost of capital parameter has a significant effect on the newsvendor fractiles, and thus inventory recommendations, it is important to calculate the newsvendor heuristic's performance over a range of plausible cost of capital values. Figure 5 illustrates the number of percentage points by which the actual inventory cost (lost sales and cost of capital) exceeds the inventory cost under

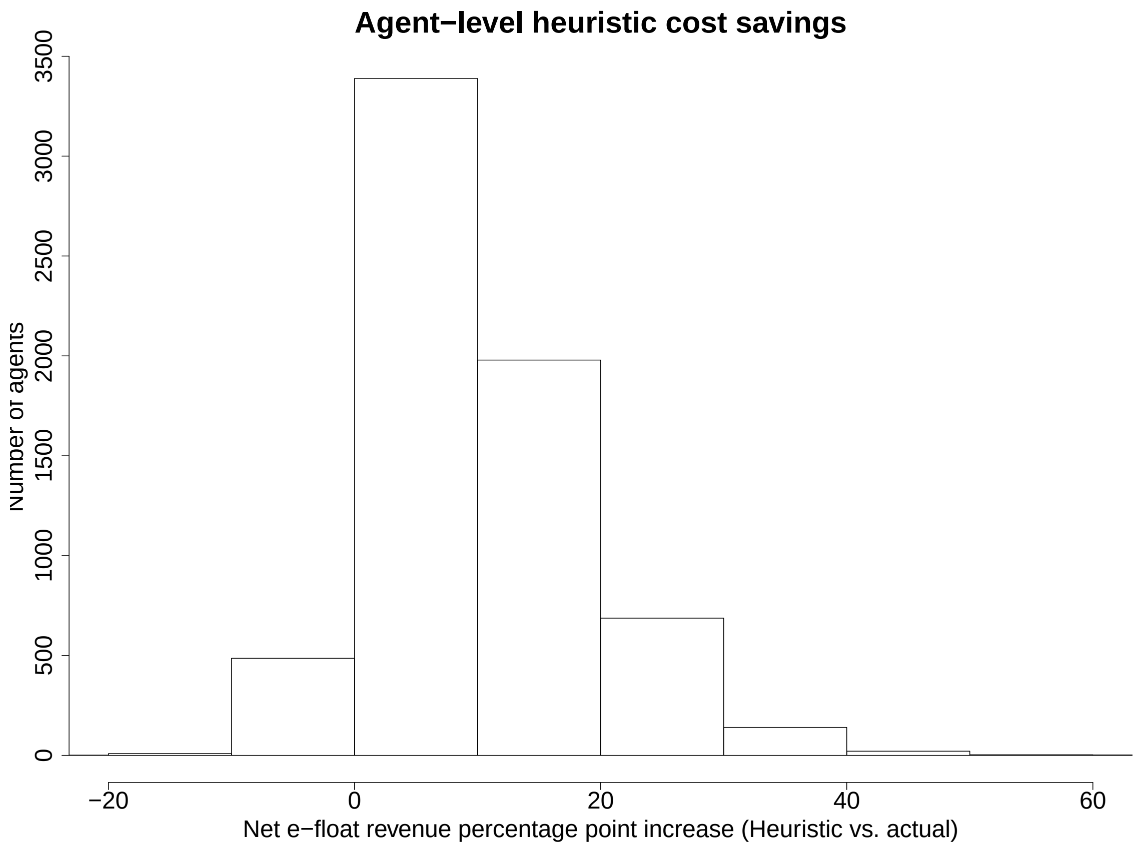

Figure 4 Agent-level e-float cost savings as a percent of total possible commission (heuristic vs. actual)

the newsvendor heuristic recommendation for various levels of cost of capital. For the range of annualized cost of capital that is likely to be applicable to the vast majority of agents (from 2% to 100%), observe that the newsvendor heuristic can increase aggregate profitability significantly, ranging from 8 to 18 percentage points of total possible revenue. Note that while the stockout loss reduction decreases as the cost of capital increases, above an annualized cost of capital of 68%, the newsvendor heuristic recommends holding less inventory than actual agent decisions in aggregate.

4.4.3. Effect of CICO balance on performance improvement From a theoretical standpoint, working capital efficiency generally is highest when agents are most balanced (the value of cash-in transactions are roughly equal to the value of cash-out transactions). In this sense, agents who are most balanced have the most to gain from recognizing the fact that sales of cash generate inventory of e-float and vice versa.

However, it seems that in general, balanced agents already take advantage of these facts. In order to illustrate this, define R as a measure of balance – the ratio of cash-in sales to total sales:

$$R = \sum_{i=1}^{N} \min(-D_i, b - q_i)^+ / (\sum_{i=1}^{N} \min(-D_i, b - q_i)^+ + \sum_{i=1}^{N} \min(D_i, q_i)^+).$$

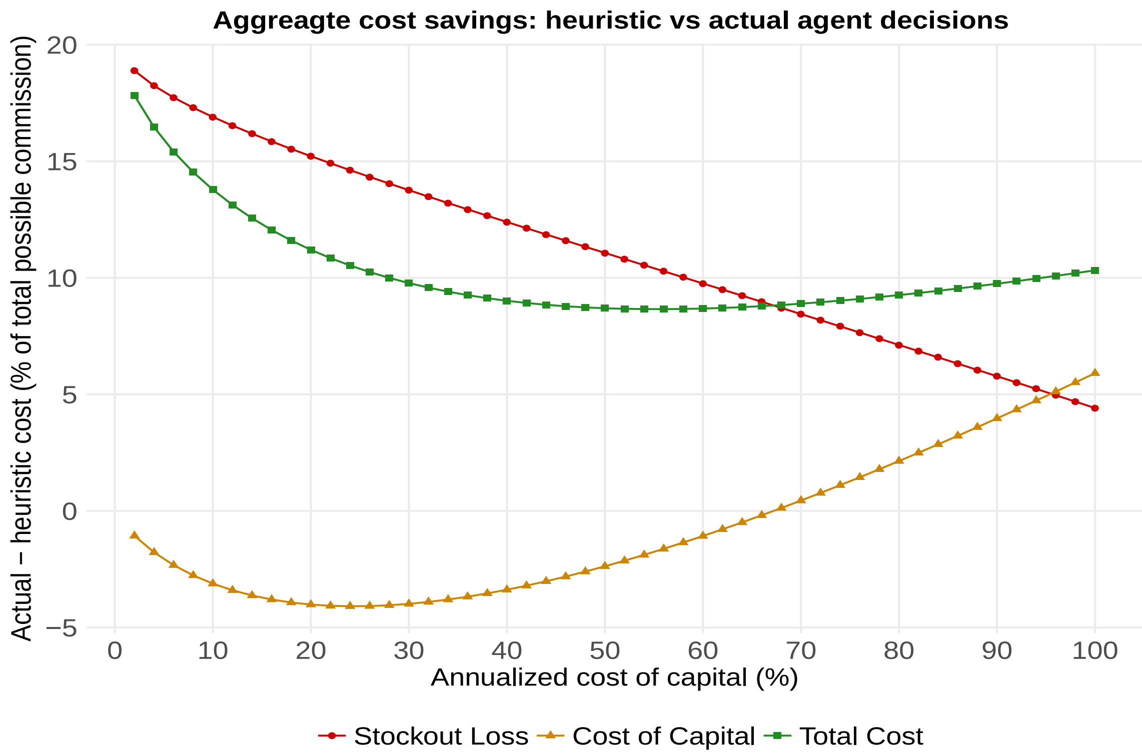

Figure 5 Cost savings sensitivity as a function of cost of capital: performance differential between actual agent decisions and newsvendor heuristic recommendations

Figure 6 illustrates agents' average cost reduction achieved through the newsvendor heuristic (as a percent of total possible e-float revenue) with respect to the ratio of e-float sales to total sales. On average, agents who are most balanced (that is, have an R value closest to 0.5) would realize the least benefit from the heuristic in terms of reduced stockout and working capital cost. On average, agents who are most imbalanced (R values close to either 0 or 1) would see the most significant gains in cost reduction. As seen in Figure 6, agents who are cash-in heavy (R close to 1) benefit the most from stockout loss reduction resulting from increased e-float inventory holding. On the other side of the spectrum, agents who are cash-out heavy (equivalently, cash-in light − R close to 0) gain the most by reducing e-float inventory (though this is coupled with a slight increase in stockout losses).

These observations are analogous to the pull-to-center effect ubiquitous in the behavioral inventory literature (e.g., Schweitzer and Cachon (2000), Bolton and Katok (2008), and Bolton et al. (2012)). In that literature, newsvendor decision-makers are shown to position quantities closer to expected demand than the optimal quantity recommends—i.e., they under-order if the optimal quantity is greater than expected demand, and over-order if the optimal quantity is less than expected demand. Similarly, results here are consistent with agents positioning their cash and e-float inventory as though they expect cash-in and cash-out demand to be more balanced (an R

closer to 0.5) than their historical demand would suggest.

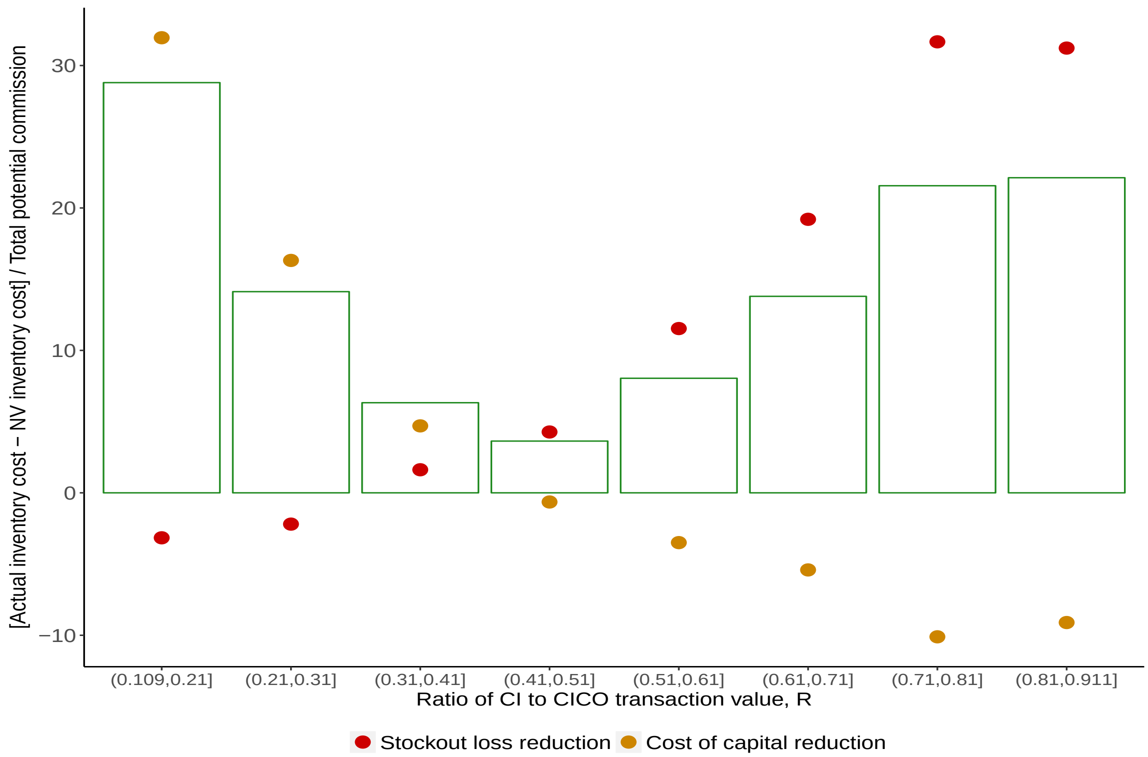

Figure 6 Performance differential between actual agent decisions and newsvendor heuristic recommendations by level of agent CICO balance

5. Discussion and Conclusion

In this section, the implications of these findings, limitations of this work, and promising areas for future mobile money research with an operations management lens are discussed.

Both the base model and the newsvendor heuristic model can produce per-agent per-day recommendations for cash and e-float stocking decisions that can increase agent profitability. While it is likely that more sophisticated estimation processes could enhance the performance of the base model, the newsvendor heuristic performs the best "out-of-the-box" and is thus well-positioned operationalized in the field given that it requires limited computational resources and comparatively minimal estimation. The newsvendor heuristic also has the added benefit of being intuitive: the concepts of minimum and maximum cumulative demand, in addition to balancing overage and underage costs, can be taught to agents. This has the potential to contribute to a virtuous cycle: less stockouts (without excess capital) lead to more profitable agents as well as happier customers.

However, there are limitations to this work. First, the estimated benefits presented here assume that agents follow the heuristic's budget and inventory recommendations. Behavioral biases that might influence how an agent interprets and acts on these recommendations are not addressed. Implementing the newsvendor heuristic in the field to determine how well agents adhere to the recommendation, as well as determining how much agents benefit in practice is thus an important area for future work.

Second, the lack of data related to cash inventory balance hampers our ability to precisely determine agent cash stocking decisions and cash stockouts. While the method employed here to calculate cash balances (and cash stockouts) is likely too conservative (likely overestimating agent cash balances and underestimating agent cash stockouts), it is likely that agents would benefit from following newsvendor heuristic cash holding recommendations in addition to e-float holding recommendations. Given that e-float stockout losses are significant, it is likely that cash stockout losses are also significant. In fact, cash stockout losses are likely to be more severe than e-float stockout losses because cash commissions are larger than e-float commissions, by approximately 50% for the average agent.

Third, it is assumed that agents are not able to rebalance during working hours. This assumption is supported by a large survey of mobile money agents in Uganda, Tanzania, and Kenya. The survey, conducted by the Helix Institute of Digital Finance, finds that the median number of rebalances (proactively converting e-float to cash or vice versa) per month in East Africa is less than 8 (McCaffrey et al. 2014, Githachuri et al. 2014). This finding, combined with the fact that in order to rebalance during working hours, agents must either close their shop or be short-staffed (to send themselves or an employee away from the shop), suggests that most agents, in general, do not rebalance during working hours. The other rebalancing assumption is that agents are able to costlessly rebalance each day (either before the first transaction and/or after the last transaction of each day). This assumption is also supported by the agent survey; the survey finds that most agents can rebalance easily and inexpensively (excluding working hour rebalances which would incur potential lost sales): 72% of agents in Uganda were within 15 minutes of a rebalancing point, and the transit cost to a rebalancing point was nominal for most agents. Conditions for rebalancing in Kenya and Tanzania were found to be similar or even more favorable (McCaffrey et al. 2014, Githachuri et al. 2014).

Last, some mobile money markets are competitive in that there are multiple, competing mobile money platforms. In some of these markets, the operators do not have the market power required to demand agent exclusivity: in these markets, most agents provide CICO services for multiple platforms. In this case, agents must make stocking decisions for each operator's platform (no scaled mobile money market has developed systems that allow simple and free exchange across platforms). However, this scenario is more complicated because, while electronic currency is not interchangeable between platforms (i.e. e-float of platform A cannot satisfy a cash-in arrival for platform B e-float), cash is fungible (i.e. the same pool of cash can be used to satisfy cash-out transactions on both platform A and B). Thus, the multi-platform agent's problem is another potential area for future research.

5.1. Conclusion

This paper introduces the context of mobile money and the mobile money agent's challenge of balancing inventory costs (expected stockout losses with cost of capital) for both cash and e-float. This setting presents a unique inventory challenge: how should a firm stock when the sales of one good generates inventory of another? A base model and an analytical heuristic that bears strong resemblance to the newsvendor model are used to generate recommendations for starting inventories of cash and e-float. These recommendations are tested against the actual decisions made by mobile money agents in an East African country. While recommendations from both models can increase agent profitability—increasing revenue net of cost of capital by 15%—the newsvendor heuristic does so while also offering substantial computational advantages relative to the base model.

Acknowledgments

Sincerest thanks to Gautier Kings, Sebastien Deletaille, Pierre Boel, Olivier Thierry, Thibault Rouby, and Maxime Temmerman of Real Impact Analytics for setting up the host country server, anonymizing data, and generally coordinating data access. Thanks also to the mobile network partner in East Africa who supported this research. Sincerest thanks also to Joel Goh and Kris Johnson Ferreira for very helpful guidance and encouragement. Thanks also to Andrew Marder of Harvard Business School's Research Computing Services for his invaluable code optimization assistance. The authors are grateful to the Harvard Business School Doctoral Programs Office for generous research support.

References

Agrawal, Mohit. 2009. Mobile Money Transfer. URL `http://www.telecomcircle.com/2009/05/mobile-money-transfer-mmt/`.

Aker, Jenny C, Rachid Boumnijel, Amanda Mcclelland, Niall Tierney. 2011. Zap It to Me : The Short-Term Impacts of a Mobile Cash Transfer Program.

Balasubramanian, Karthik, David Drake. 2015. Mobile money: The effect of service quality and competition on demand .

Batista, Cátia, Pedro C Vicente. 2013. Introducing Mobile Money in Rural Mozambique: Evidence from a Field Experiment.

BCG. 2011. The Socio-Economic Impact of Mobile Financial Services. Tech. Rep. April.

Bolton, Gary E, Elena Katok. 2008. Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management* **10**(3) 519–538.

Bolton, Gary E, Axel Ockenfels, Ulrich W Thonemann. 2012. Managers and students as newsvendors. *Management Science* **58**(12) 2225–2233.

BTCA. 2014. Better Than Cash Alliance: for the Development Community. URL `http://betterthancash.org/join/for-the-developmentc-ommunity/`.

Chen, Xin, David Simchi-Levi. 2009. a New Approach for the Stochastic Cash Balance Problem With Fixed Costs. *Probability in the Engineering and Informational Sciences* **23**(04) 545.

Chod, Jiri, Jianer Zhou. 2013. Resource flexibility and capital structure. *Management Science* **60**(3) 708–729.

Edgeworth, Francis Y. 1888. The mathematical theory of banking. *Journal of the Royal Statistical Society* **51**(1) 113–127.

Fleischmann, Moritz, Jacqueline M. Bloemhof-Ruwaard, Rommert Dekker, Erwin van der Laan, Jo a.E.E. van Nunen, Luk N. Van Wassenhove. 1997. Quantitative models for reverse logistics: A review. *European Journal of Operational Research* **103**(1) 1–17.

Girgis, Nadia. 1968. Management Science, **15**(3) 130–140.

Githachuri, K., M. McCaffrey, L. Anthony, A. Schiff, A.M. van Swinderen, G. A. N. Wright. 2014. Agent network accelerator survey :Tanzania country report 2013. Tech. rep., The Helix Institute of Digital Finance.

Groupe Speciale Mobile Association. 2015. State of the Industry: Mobile Financial Services for the Unbanked. Tech. rep.

Harrison, J Michael, Allison J Taylor. 1978. Optimal control of a brownian storage system. *Stochastic Processes and Their Applications* **6**(2) 179–194.

Hausman, Warren H, Antonio Sanchez-bell. 1975. The Sochastic Cash Problem with Average Compensating-Balance Requirements. *Management Science* **21**(8) 849–857.

Helms, Brigit. 2006. Access for All. Tech. rep.

Intermedia. 2013. Mobile Money Use, Barriers, and Opportunities. Tech. Rep. February.

Jack, William, Tavneet Suri. 2014. Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution . *American Economic Review* **104**(1) 183–223.

Kendall, Jake. 2011. An Emerging Platform: From Money Transfer System to Mobile Money Ecosystem.

Khouja, Moutaz. 1999. The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega* **27**(5) 537–553.

Mas, Ignacio. 2010. Savings for the Poor. *World Economics* **11**(4) 1–12.

Mbiti, Isaac. 2011. Mobile Banking: the Impact of M-Pesa in Kenya.

Mbiti, Isaac, David N Weil. 2013. The Home Economics of E-Money: Velocity, Cash Management, and Discount Rates of M-Pesa Users. *American Economic Review* **103**(3) 369–374.

McCaffrey, M., L. Anthony, A. Schiff, K. Githachuri, G. A. N. Wright. 2014. Agent Network Accelerator Survey: Kenya Country Report. Tech. rep., The Helix Institute of Digital Finance.

Morawczynski, Olga. 2009. Poor People Using Mobile Financial Services: Observations on Customer Usage and Impact from M-PESA.

Neave, Edwin H, Source Management Science, Theory Series Mar. 1970. The Stochastic Cash Balance Problem with Fixed Costs for Increases and Decreases. *Management Science* **16**(7) 472–490.

Netessine, Serguei, Fuqiang Zhang. 2005. Positive vs. negative externalities in inventory management: Implications for supply chain design. *Manufacturing & Service Operations Management* **7**(1) 58–73.

Porteus, Evan L, Edwin H Neave. 1972. Stochastic Cash Balance Problem with Charges Levied against the Balance. *Management Science* **18**(11) 600–602.

Schweitzer, Maurice E, Gérard P Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* **46**(3) 404–420.

Silver, Edward Allen, David F Pyke, Rein Peterson, et al. 1998. *Inventory management and production planning and scheduling*, vol. 3. Wiley New York.

Suri, Tavneet, William Jack. 2016. The long-run poverty and gender impacts of mobile money. *Science* **354**(6317) 1288–1292.

Van Mieghem, Jan A. 2007. Risk mitigation in newsvendor networks: Resource diversification, flexibility, sharing, and hedging. *Management Science* **53**(8) 1269–1288.

van Ryzin, Garrett J, Kalyan T Talluri. 2005. An introduction to revenue management. *Emerging Theory, Methods, and Applications*. INFORMS, 142–194.

Appendix

A. Reflected Brownian motion heuristic

A.1. Description

An agent's inventory problem can also be approached as a stochastic flow problem with a fixed buffer size. As in most stochastic flow settings, flow (in this case, CICO demand) can be positive or negative. Furthermore, an agent loses cash sales when the buffer is empty (i.e. cash quantity is equal to zero) and e-float sales when the buffer is full (i.e. when cash quantity is equal to the budget, and thus e-float quantity is zero). Given characteristics of an agent's demand, as well as parameters m_c, m_e, and γ, the goal is to again determine both the optimal budget (buffer size) b^* and the optimal starting cash inventory q_1^*. The optimal e-float quantity again follows directly as

$b^* - q_1^*$. By assuming demand can be approximated as a series of i.i.d. normally-distributed random variables (within an agent-day), each with mean μ_D and standard deviation σ_D, the inventory process can be approximated as a RBM. Let net demand Δ_t be a (μ, σ) Brownian motion with parameters μ_D and σ_D respectively. Next, define q_t as the RBM representing the agent's cash inventory position, with initial cash inventory represented as q_1. Further, define upper and lower control processes U_t and L_t invoked when $q_t = b$ and $q_t = 0$, respectively. These upper and lower control processes force the quantity of cash q_t to stay within the bounds of 0 and b, respectively. Thus, the cash quantity at arrival t can be expressed as

$$q_t = q_1 - \Delta_t + L_t - U_t.$$

Written in this way, it is clear that the value of U_t represents the cumulative magnitude of lost e-float sales up to arrival t and the value of L_t represents the cumulative magnitude of lost cash sales up to arrival t. Both L_t and U_t are continuous and increasing processes with $L_0 = U_0 = 0$. Thus, the underage cost after t arrivals is $m_c \cdot L_t + m_e \cdot U_t$.

Note that the actual agent's problem features a discrete number of arrivals – and can be seen as an RBM sampled at integer time-points. It follows that the underage cost captured by the upper and lower control processes $(m_c \cdot L_t + m_e \cdot U_t)$ represents an upper bound on the underage cost of the base model; the intuition here is that the RBM can accumulate costs between these integer time-points (discrete demand arrivals) while the actual agent problem only realizes underage costs at these discrete demand arrival time-points. Accordingly, the RBM heuristic provides a conservative approximation of underage cost.

Proposition 4 *The underage of the reflected Brownian motion model is an upper bound on the underage of the base model with normally distributed demand increments*

To build out the heuristic, we again assume that the number of arrivals is a geometric random variable with success parameter λ, where λ is the probability that any given arrival will be the final arrival. The expected discounted cost function (of underage, in this case) of an RBM with two control barriers (at 0 and b, respectively) can be written as follows:

$$k(q_1) = E\left[\int_0^\infty e^{-\lambda t}\left[m_c dL + m_e dU\right]\right]$$

Harrison and Taylor (1978) show that by applying Ito's lemma, finding $k(q_1)$ requires solving only the ordinary differential equation:

$$\frac{1}{2}\sigma^2 k''(q_1) + \mu k'(q_1) - \lambda k(q_1) = 0,$$

where $0 \leq q_1 \leq b$ and the boundary conditions are:

$$k'(0) = -m_c \ \text{ and } \ k'(b) = m_e.$$

Solving this ordinary differential equation results in the underage cost function $k(q_1)$. This cost function (too unwieldy to present here) is convex in q_1, allowing for closed form characterizations of q_1^* and $k(q_1^*)$ which are both also too unwieldy to present here. With q_1^* and $k(q_1^*)$, the optimal budget b can be found by minimizing the total cost function $G(\cdot)$, which is the sum of the underage cost and the cost of capital.

$$\min_b G(q_1, b) = k(q_1^*, b) + \gamma b$$

While the complexity of this function precludes a closed-form analytical solution, numerical optimization can generate b^* given parameters μ_D, σ_D, λ, m_e, m_c, and γ. Recommendations can now be generated: the optimal amount of cash to stock daily is q_1^* and the optimal amount of e-float stock daily is $b^* - q_1^*$. However, note that the RBM heuristic's results are only valid to the extent that the assumption of normal demand holds.

A.2. Performance and limitations

While the RBM heuristic does significantly outperform actual agent decisions, it under-performs both the base model and the newsvendor heuristic. While the RBM heuristic does feature lower stockout losses than the newsvendor heuristic (3.3% to 5.9%), it requires significantly more working capital (16.4% to 12.3%). The same is true when comparing the RBM heuristic with the base model, but to a slightly lesser degree: the RBM heuristic results in lower stockout losses (3.3% to 4.2%) and higher working capital requirements (16.4% to 14.%). This observation is a direct result of the fact that the RBM heuristic is built upon an upper bound on underage cost – which leads to higher inventory recommendation. This higher inventory recommendation results in both lower stockout losses and greater capital costs as illustrated in Figure A.1 below.

There are two reasons that the RBM heuristic underperformcs the NV heuristic. First, the normal approximation of the empirical demand distribution for each agent each day does not fit the actual distribution of demand in this context. Specifically, it is the case that the empirical demand distribution can be characterized by higher kurtosis (fatter tails) than the normal distribution's kurtosis of 3. The mean kurtosis of empirical demand distributions for agents in the sample was 42. The minimum kurtosis of the empirical demand distributions is 3.2, implying every agent's distribution has fatter tails than the normal distribution. The normal distribution approximation may be underestimating the probability of more extreme events. Thus, If the empirical demand

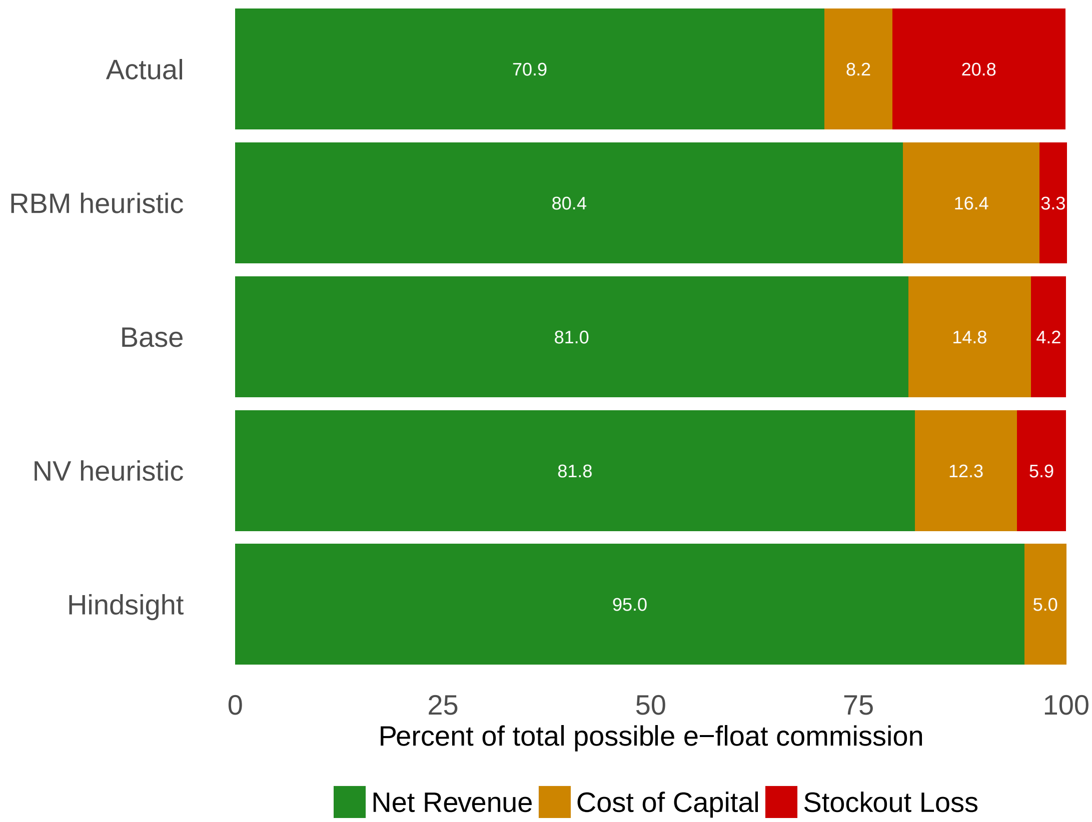

Figure A.1 Performance evaluation: aggregate net e-float revenue for actual agent decisions, RBM heuristic, base model, newsvendor heuristic, and hindsight recommendations over 180 days

distributions were more normal, the stockout losses associated with the RBM heuristic might be even lower. Second, the RBM heuristic, like the base model, requires two sets of separate inputs (estimates for both number of arrivals and also distribution of magnitudes of demand); the compounding of estimation errors may be hampering performance. The newsvendor heuristic, on the other hand, requires only a single set of inputs (the distributions of the maximum or minimum cumulative demand).

B. Proofs

Proof of Lemma 1: It is sufficient to show that the cash and e-float underage, as well as the inventory evolution can be written in terms of D_t. Let $D_t = D_t^c - D_t^e = \mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^e>0} \cdot D_t^e$.

$$\begin{aligned}
Underage &= (D_t^c - q_t)^+ + (D_t^e - f_t)^+ \\
&= \mathbb{1}_{D_t^c>0}(D_t^c - q_t)^+ + \mathbb{1}_{D_t^e>0}(D_t^e - f_t)^+ \\
&= (\mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (\mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^e>0} \cdot f_t)^+ \\
&= (\mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (-\mathbb{1}_{D_t^c>0} \cdot D_t^c + \mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^e>0} \cdot f_t)^+
\end{aligned}$$

$$= (D_t - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (-D_t - \mathbb{1}_{D_t^e>0} \cdot f_t)^+$$
$$= (D_t - q_t)^+ + (-D_t - b + q_t)^+$$

For inventory evolution:

$$\begin{aligned} q_{t+1} &= \mathbb{1}_{D_t^e>0} \cdot \min(b, q_t + D_t^e) + \mathbb{1}_{D_t^c>0} \cdot \max(q_t - D_t^c, 0) \\ &= \min(b, \max((q_t - \mathbb{1}_{D_t^c>0} \cdot D_t^c + \mathbb{1}_{D_t^e>0} \cdot D_t^e), 0) \\ &= \min(b, \max(q_t - D_t, 0)) \quad \square \end{aligned}$$

Proof of Proposition 1.A: The inventory evolution equation is:

$$q_{i+1} = q_i - \mathbb{1}_{D_i>0} \cdot \min(q_i, D_i) + \mathbb{1}_{D_i<0} \cdot \min(b - q_i, -D_i)$$

By the definition of $\hat{\Delta}$ and $\check{\Delta}$, as well as $q_1 \geq \hat{\Delta}$ and $(b - q_1) \geq -\check{\Delta}$

$$q_1 \geq \max_{1 \leq t \leq N} (\sum_{i=1}^{t} D_i) \Rightarrow q_1 \geq D_1$$
$$b - q_1 \geq -\min_{1 \leq t \leq N} (\sum_{i=1}^{i} D_i) \Rightarrow b - q_1 \geq -D_1$$

Substituting into the inventory evolution equation:

$$\begin{aligned} &q_2 = q_1 - D_1 \\ &q_3 = q_2 - \mathbb{1}_{D_2>0} \cdot \min(q_2, D_2) + \mathbb{1}_{D_2<0} \cdot \min(b - q_2, -D_2) \\ &\quad = q_1 - D_1 - \mathbb{1}_{D_2>0} \cdot \min(q_1 - D_1, D_2) + \mathbb{1}_{D_2<0} \cdot \min(b - q_1 + D_1, -D_2) \end{aligned}$$

Using the fact that $q_1 \geq D_1 + D_2$ and $b - q_1 \geq -(D_1 + D_2)$:

$$q_3 = q_1 - D_1 - D_2$$
$$\ldots$$
$$q_N = q_1 - \sum_{i=1}^{N-1} D_i$$

Then:

$$q_i - D_i \geq 0 \ \forall \ i \leq N$$
$$b - q_i + D_i \geq 0 \ \forall \ i \leq N$$

It follows immediately that:

$$m_c \sum_{i=1}^{N} (D_i - q_i)^+ + m_e \sum_{i=1}^{N} (-D_i - (b - q_i))^+ = 0 \ \square$$

Proof of Proposition 1.B: Starting from proposition 1.A:

$$(\hat{\Delta}_t)^+(-\check{\Delta}_t)^+ = (\max_{1\leq j\leq t}\sum_{i=1}^{j} D_i)^+ + (-\min_{1\leq j\leq t}\sum_{i=1}^{j} D_i)^+$$
$$< \sum_{i=1}^{t}(D_i)^+ + \sum_{i=1}^{t}(-D_i)^+ \quad \square$$

Proof of Proposition 2.A: Let l_t and u_t represent the cumulative underage cost of cash and e-float respectively of the base model after t arrivals. These values can then be compared to the newsvenduer heuristic underage to directly prove the result. For cash underage:

$$\begin{aligned} l_t &= \max_{1\leq j\leq t}(q_1 - \Delta_j - u_j)^- \\ &= \max_{1\leq j\leq t}(\Delta_j - q_1 + u_j)^+ \\ &\geq \max_{1\leq j\leq t}(\Delta_j - q_1)^+ \\ &= (\hat{\Delta} - q_1)^+ \end{aligned}$$

Analogously, for e-float underage, the result follows from:

$$\begin{aligned} u_t &= \max_{1\leq j\leq t}(b - q_1 + \Delta_j - l_j)^- \\ &= \max_{1\leq j\leq t}(-\Delta_j + l_j - b + q_1)^+ \\ &\geq \max_{1\leq j\leq t}(-\Delta_j - b + q_1)^+ \\ &= (-\check{\Delta}_t - b + q_1)^+ \end{aligned}$$

Proof of Proposition 2.B: There are four cases: Case A: $l_t > 0$ and $u_t = 0$, Case B: $l_t = 0$ and $u_t > 0$, Case C: $l_t > 0$ and $u_t > 0$, and Case D: $l_t = 0$ and $u_t = 0$. Note that when there is no base model cash and/or e-float underage (i.e. $l_t = 0$ and/or $u_t = 0$), sharpness follows from Proposition 2.A and the fact that heuristic underage is non-negative. This demonstrates sharpness in Case D. Sharpness is further demonstrated for positive underage in Cases A and B, while non-sharpness is demonstrated in Case C.

$$\textbf{Case A: } \begin{aligned} l_t &= \max_{1\leq j\leq t}(q_1 - \Delta_j - u_j)^- \\ &= \max_{1\leq j\leq t}(q_1 - \Delta_j - 0)^- \\ &= \max_{1\leq j\leq t}(q_1 - \Delta_j)^- \end{aligned}$$

$$\textbf{Case B: } \begin{aligned} u_t &= \max_{1\leq j\leq t}(b - q_1 + \Delta_j - l_j)^- \\ &= \max_{1\leq j\leq t}(-\Delta_j + 0 - b + q_1)^+ \\ &= (-\check{\Delta}_t - b + q_1)^+ \end{aligned}$$

$$\textbf{Case C: } l_t = \max_{1\le j\le t}(q_1-\Delta_j-u_j)^- \\ > \max_{1\le j\le t}(q_1-\Delta_j)^-$$

$$u_t = \max_{1\le j\le t}(b-q_1+\Delta_j-l_j)^- \\ > (-\check{\Delta}_t-b+q_1)^+ \quad \square$$

Proof of convexity of equation 2 The cost function used to derive the newsvendor heuristic:

$$G(q_1,b)=\mathbb{E}[(m_c\cdot(\hat{\Delta}-q_1)^+]+\mathbb{E}[m_e\cdot(-\check{\Delta}-(b-q_1))^+]+\gamma\cdot b)$$

The function's hessian matrix is positive-semidefinite:

$$\mathbf{H}G(q_1,b)=\begin{bmatrix}\frac{\partial^2 G(q_1,b)}{\partial q_1^2} & \frac{\partial^2 G(q_1,b)}{\partial q_1\partial b}\\ \frac{\partial^2 G(q_1,b)}{\partial b\partial q_1} & \frac{\partial^2 G(q_1,b)}{\partial b^2}\end{bmatrix}$$

To show that this matrix is positive semi-definite, it is sufficient to show that the Hessian's principal minors are non-negative: $\frac{\partial^2 G(q_1,b)}{\partial b^2}\ge 0$, $\frac{\partial^2 G(q_1,b)}{\partial q_1^2}\ge 0$, and $\frac{\partial^2 G(q_1,b)}{\partial b^2}\frac{\partial^2 G(q_1,b)}{\partial q_1^2}-(\frac{\partial^2 G(q_1,b)}{\partial b\partial q_1})^2\ge 0$. m_c, m_e and probabilities are non-negative, so

$$\frac{\partial^2 G(q_1,b)}{\partial q_1^2}=m_c\cdot f_{\hat{\Delta}}(q_1)+m_e\cdot f_{\check{\Delta}}(q_1-b)\ge 0,$$
$$\frac{\partial^2 G(q_1,b)}{\partial b^2}=me\cdot f_{\check{\Delta}}(q_1-b)\ge 0,$$
$$\frac{\partial^2 G(q_1,b)}{\partial b^2}\frac{\partial^2 G(q_1,b)}{\partial q_1^2}-(\frac{\partial^2 G(q_1,b)}{\partial b\partial q_1})^2=m_c\cdot f_{\hat{\Delta}}(q_1)\cdot m_e\cdot f_{\check{\Delta}}(q_1-b)\ge 0. \quad \square$$

Proof of Proposition 3: The first order condition of equation 2 with respect to q_1 yields the optimality condition:

$$m_c\cdot(1-F_{\hat{\Delta}}\,(q_1))=m_e\cdot(F_{\check{\Delta}}\,(q_1-b)) \tag{3}$$

The first order condition with respect to b results in:

$$F_{\check{\Delta}}(q_1-b)=\frac{\gamma}{m_e} \tag{4}$$

Substituting 4 into 3 yields:

$$F_{\hat{\Delta}}(q_1)=1-\frac{\gamma}{m_c}$$

Thus, the optimal q_1 and f_1 are:

$$q_1^*=\left(F_{\hat{\Delta}}^{-1}\left(1-\frac{\gamma}{m_c}\right)\right)^+$$
$$f_1^*=\left(-F_{\check{\Delta}}^{-1}\left(\frac{\gamma}{m_e}\right)\right)^+ \square$$

Proof of Proposition 4: Let the cumulative underage cost of cash and the underage cost of e-float in the base model be represented as l_t and u_t respectively after t arrivals. Therefore, it is sufficient to show that $U_t \geq u_t$ and $L_t \geq l_t$ $\forall t \in (0,1,2,...)$. Proceeding with a proof by induction, the base case holds:

$$L_0 = l_0 = 0$$
$$U_0 = u_0 = 0$$

Let the inductive hypothesis be:

$$L_j \geq l_j \text{ and } U_j \geq u_j \ \forall \ j \in (0,1,...,t-1)$$

Each arrival (and thus underage) is either e-float or cash, thus $\Delta u_t \cdot \Delta l_t = 0$. There are thus three cases: Case A: $\Delta u_t = 0$ and $\Delta l_t > 0$, Case B: $\Delta l_t = 0$ and $\Delta u_t > 0$and Case C: $\Delta u_t = \Delta l_t = 0$. Let $X_t = q_1 - \Delta_t$ and x_t be a sample of X_t at integer time-points.

$$\begin{aligned}
\textbf{Case A: } U_t &= \sup_{0\leq s\leq t} (b - X_s - L_s)^- \\
&\geq \sup_{s\in(0,...,t)} (b - X_s - L_s)^- \\
&\geq \sup_{s\in(0,...,t-1)} (b - X_s - L_s)^- \\
&\geq \sup_{s\in(0,...,t-1)} (b - x_s - l_s)^- && \text{by inductive hypothesis} \\
&= u_{t-1} \\
&= u_t && (1)
\end{aligned}$$

$$\begin{aligned}
L_t &= \sup_{0\leq s\leq t} (X_s - U_s)^- \\
&\geq \sup_{s\in(0,...,t)} (X_s - U_s)^- \\
&\geq \sup_{s\in(0,...,t-1)} (x_s - u_s)^- && \text{by (1)} \\
&= l_t
\end{aligned}$$

$$\begin{aligned}
\textbf{Case B: } L_t &= \sup_{0\leq s\leq t} (X_s - U_s)^- \\
&= \sup_{0\leq s\leq t-1} (X_s - U_s)^- \\
&\geq \sup_{s\in(0,...,t-1)} (X_s - U_s)^- && \text{by inductive hypothesis} \\
&\geq \sup_{s\in(0,...,t-1)} (x_s - u_s)^- \\
&= l_{t-1} \\
&= l_t && (2)
\end{aligned}$$

$$
\begin{aligned}
U_t &= \sup_{0\le s\le t}(b - X_s - L_s)^- \\
&\ge \sup_{s\in(0,\dots,t)}(b - X_s - L_s)^- \\
&\ge \sup_{s\in(0,\dots,t-1)}(b - x_s - l_s)^- && \text{by (2)} \\
&= u_t
\end{aligned}
$$

Case C follows directly from Case A and Case B. $\square$