Tactics for approaching cash optimisation in bank branches

Miguel Aguilar Zaragoza* and Idalia Flores de la Mota

Universidad Nacional Autónoma de México,
Av. Universidad N° 3000, C.U., Del. Coyoacán,
Ciudad de México, DF. CP 04510, México
Email: maguilarz@live.com
Email: idalia@unam.mx
*Corresponding author

Abstract: Banks, through their branches, offer cash transactions as a key service, so any lack of available cash is a critical issue, since it affects the prestige of the entire bank not just of that branch in particular. This paper aims to establish a *vault policy*, that lets the staff know how to manage the branch's money efficiently, by handling orders to the central vaults with realistic assumptions and easy-to-implement criteria. We present a model based on dynamic programming principles, to represent the problem and generate an input for the vault policy. Other inputs for the policy are some parameters set according to a branch's demand for a cash transactions approach, from the perspective that the best fit is between frequentist and Bayesian. With these inputs and a definition of some 'intuitive' rules we can implement and assess the vault policy. The results for 30 branches are presented.

Keywords: dynamic programming; compound process; generalised linear models; GLMs; integrated nested Laplace approximation; INLA; R; cash optimisation; bank branches.

Reference to this paper should be made as follows: Zaragoza, M.A. and Flores de la Mota, I. (2016) 'Tactics for approaching cash optimisation in bank branches', *Int. J. Simulation and Process Modelling*, Vol. 11, No. 6, pp.492–503.

Biographical notes: Miguel Aguilar Zaragoza holds a BS in Actuarial Sciences from ITAM, and is currently studying his Masters in Operations Research at UNAM. He has worked in the public and private sectors in many activities connected with optimisation/ongoing improvement. Nowadays, he is working in an international firm in the area responsible for its operations strategy in branches. Statistic inference and stochastic processes are the main topics he is passionately interested in.

Idalia Flores de la Mota is a Mathematician from the Faculty of Sciences of the UNAM. She studied her Masters and PhD in Operations Research in the Faculty of Engineering of the UNAM. She has published notes, chapters in books, booklets and articles for international journals. She has been a referee in the following journals: *Simulation*; the *Journal of Accounting and Administration*; *Computing Reviews Online* and *Revista Iberoamericana de Automática e Informática Industrial*. She belongs to the Institute for Operations Research and the Management Science and she is the Director of the Center for Simulation McLeod in Mexico.

This paper is a revised and expanded version of a paper entitled 'Approaching demand of cash transactions at bank branches' presented at EMSS 2014, Multi-purpose Complex Agora, University of Bordeaux, Bordeaux, France, 10–12 September 2014.

1 Introduction

Within their administration, banks have the ongoing task of making their activities more efficient, i.e., meeting service levels according to customer expectations at a minimal cost. Among their main activities, banks offer services through their tellers and automated devices that generate cash inflows and outflows in their branches.

The purpose of looking for more efficient cash management in bank branches is to try to minimise money-transfer cost, opportunity cost (caused by the storage of money in the vaults of the branches) and the amount exposed to the risk of theft, while maintaining the operation of these entities at a desired level. All these decisions are better supported by a truthful approach to cash transactions. Since it is the most representative part of cash services, and therefore 'more predictable', this article only covers local currency, so it does not include foreign currencies.

Thus, our objective is to propose a *vault policy* that includes order recommendations based on a model that is of assistance in finding out how much money is needed in each branch of a bank and thus make it easier for the person in charge of the branch vault to manage it, with the support of some rules.

Simutis et al. (2007) addressed the cash management issue for ATMs, presenting the complexity of the best known commercial solutions at the time, which, by the way, also cover cash solutions for branch vaults. We agree with his diagnosis of their inefficiency, result of the models' paradigm, which affects development time as well as ongoing maintenance. Therefore, these commercial solutions do not properly meet banks' needs, making this a big area of opportunity that we still need to address.

Cabello (2013) has written about efficient cash management in bank branches. Her model is very attractive because it is easy to implement. It seems to have more efficiency gains in small and medium sized branches as well as in those branches where there are lots of fluctuations, e.g., active urban areas or cash centres. However, there are no completely realistic assumptions, and this can cause solutions with unexpected consequences, including negative *stop costs* (transfer costs).

Some people have also published proposals for the management of foreign currencies in bank branches. In his paper, Bell and Hamidi-Noori (1984) discuss this issue as one of *inventory management* and the results are found by using decision rules for any bank branch with a considerable volume of foreign exchange transactions. The idea of treating vault stocks as inventory is indeed an attractive one. However, foreign exchange demand is a very particular problem and, as we have seen, these rules are not easy to implement.

Furthermore, some proposals have been made to determine the optimal stock levels that would ensure that the cash is available to cover the future needs of any entity. Girgis (1968) proposes an optimal inventory policy solely for the case where the costs of storing or not storing money are convex functions, in a scenario where the decision to increase or decrease inventory levels has no fixed cost. On the other hand, Chen (1991) addresses the problem of the amount of money a bank should maintain by formulating a *dynamic programming* model that assumes: independent and identically distributed demands every week, only the case of positive demands and no fixed costs for immobilisation or transfers. This model does not determine any maximum stored amount and tolerates a level of shortage as well, both undesirable aspects.

All the above papers have addressed their problems giving little attention to modelling demand with any significant degree of accuracy. As we are suggesting a model that uses some parameters set up with the projected demand for cash transactions, we need to approach it with a greater degree of accuracy. It is also worth mentioning that we are making assumptions that simplify what we have seen in real life, having experienced this problem close up, not just as in theory. In general, we are sure both considerations add value to previous studies.

The context of the problem is described in the following section.

2 Description of the problem

We addressed the *cash management at bank branches* problem considering the situations contained in this section.

2.1 Context

Branch managers must handle the cash flow of the vaults according to Figure 1.

Figure 1 Cash inflows/outflows in vaults of bank branches (see online version for colours)

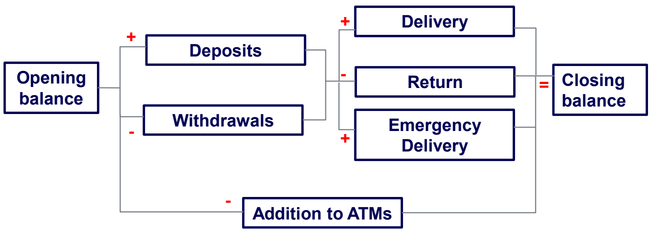


Figure 1 shows the possible movements in a branch's cash stock. The *opening balance* is the amount of the stock at the beginning of each day; *deposits* and *withdrawals* are the amounts of transactions with users in tellers; *delivery* and *emergency delivery* are cash supplies from central vaults to the branch, with different degrees of urgency; *return* is the shipment of cash from the branch to the central vaults; *addition to ATMs* represents the amount of cash coming out of the branch vault to feed their ATMs. Finally, the *closing balance* is the amount of stock at the end of the day, obtained by arithmetic operations using the prior concepts.

Traditional cash management control mechanisms in banks allow branches to have a high level of discretion in making decisions as regards ordering *deliveries* and *returns* to central vaults. Those mechanisms are based on setting bounds on the *maximum number of transfers* and the *minimum amount per transfer*, both of which bounds are required for cost purposes; and bounds on *maximum cash storage*, mainly for security purposes. Assuming that the first two bounds permit an efficient operation, it is easy to see that the third one has the power to affect the best approach for the other two parameters. Also, it is easy to see that the cost-based and security-based bounds might conflict. That is why it is important for the *maximum number of transfers* and *minimum amount per transfer* to be set once the *maximum cash storage* has been established. Even in our case (which avoids these cost-control parameters as instructions for branches), it is very important to define a suitable maximum level of cash storage that permits efficiency while respecting the security parameter, besides a *safety stock* to give confidence about the availability of money.

2.2 A dynamic programming perspective

Because of the features of the problem, particularly the dependence of current decisions on previous decisions, a *dynamic programming* approach is suggested.

Dynamic programming was developed by Richard Bellman in the 1950s. This method solves some optimisation problems by splitting them into subproblems called *stages*, while each of these *stages* has different *states*. Each subproblem is solved in a recursive sequence given by the recursive equations, so each result is an input for the subsequent subproblem. Thus, in the last *stage* the whole problem is solved.

The statement that best describes the idea of the *dynamic programming* method is its *principle of optimality*, which was determined by Bellman in 1957 and says that: "each subpolicy of an optimum policy must itself be an optimum policy with regard to the initial and the terminal states of the subpolicy" (Lin, 2002). Thus, an optimisation problem can be addressed by using *dynamic programming*, if the solution to the problem can be reached through a sequence of decisions, one in each *stage*, and that sequence satisfies the *principle of optimality*. Of course, if possible, it is necessary to define the potential *states* of the variable and the *stages* in the problem.

2.2.1 Definition of states and stages

The partition of the original problem is done by defining *stages* that could represent many things depending on the problem, such as periods of time or intervals, or steps in a procedure. At each stage, the selection of an *alternative* results in a possible state. In every *stage*, the best of these *alternatives* is chosen for each *state*. And more important suboptimal solutions are found in each *stage*.

The proposal is that the amount of cash stored in a branch vault is bounded by two concepts: *safety stock* and *maximum cash storage*, determined by the demand approach and some key feedbacks. A proposal for setting these bounds is presented in Section 2.3.

For the definition of the problem as one of *dynamic programming*, the following *states* and *stages* are stated:

- *State j*, *j* represents a possible amount of stock, defined as interval. The bounds of the interval are defined as *safety stock* and *maximum cash storage* as extreme values.
- *Stage t*, *t* = 1, 2, …, |*n*|. (|*n*|: days considered).

2.2.2 Model variables and solution

Given the previous definitions, the problem is described as one of *mixed integer linear programming* (*MILP*), and then it is adapted as a *dynamic programming problem* (*DPP*).

Let the variables be denoted by:

Del_t *delivery* on day t.

Ret_t *return* on day t.

$IDel_t$ indicator for *delivery* on day t.

$IRet_t$ indicator for *return* on day t.

Let the *parameters* be denoted by:

X_t estimated *closing balance* for day t

Dep_t estimated *deposits* for day t

Wth_t estimated *withdrawals* for day t

$AdiATM_t$ estimated *addition* to ATMs for day t

SS *safety stock*, i.e., minimum limit of stocks

MS *maximum cash storage*.

i daily cash immobilisation cost, including opportunity cost and risk fee

KD fixed cost per *delivery*

KR fixed cost per *return*

CD variable cost per *delivery*

CR variable cost per *return*

MM arbitrary great value.

The problem formulated as MILP is shown below:

Minimise:

$$\sum_{t=1}^{|n|}\Big[\big(X_{t+1}+Del_t-Ret_t+Dep_t-Wth_t -AdiATM_t\big)*+IDEL_t*KD+IRet_t*KR+Del_t*CD +Ret_t*CR\Big],$$

subject to:

$$SS\le X_{t-1}+Del_t-Ret_t+Dep_t-Wth_t-AdiATM_t\le MS \quad t=1,\dots,|n|$$

$$MM*IDel_t\ge Del_t \qquad t=1,\dots,|n|$$

$$MM*IRet_t\ge Ret_t \qquad t=1,\dots,|n|$$

$$IDel_t\in\{0,1\} \qquad t=1,\dots,|n|$$

$$IRet_t\in\{0,1\} \qquad t=1,..,|n|$$

$$IRet_t\le 1-IDel_t \qquad t=1,\dots,|n|$$

$$Del_t\ge 0 \qquad t=1,\dots,|n|$$

$$Ret_t\ge 0 \qquad t=1,\dots,|n|$$

Decisions made one day affect the cash balances of the following day, so the same problem can be posed with the explicit omission of X_{t-1} (t = 2, …, |n|), and it is only necessary to include the first day *opening balance*, which is the *closing balance* of the previous day, X_0; and, since this value is constant at each *stage*, as its cost, the problem can be represented as follows:

Minimise:[1]

$$\sum_{t=1}^{|n|}\Big\{Del_t*\big[(|n|-t+1)*i+CD\big] +Ret_t*\big[CR-(|n|-t+1)*i\big]+IDel_t*KD+IRet_t*KR\Big\},$$

subject to:

$$SS \le X_0 + \sum_{j=1}^{t} \left(Del_j - Ret_j + Dep_j - Wth_j - AdiATM_j \right) \le MS \qquad t = 1, ..., |n|$$

$$MM * IDel_t \ge Del_t \qquad t = 1, ..., |n|$$

$$MM * IRel_t \ge Ret_t \qquad t = 1, ..., |n|$$

$$IDel_t \in \{0, 1\} \qquad t = 1, ..., |n|$$

$$IRet_t \in \{0, 1\} \qquad t = 1, ..., |n|$$

$$IRet_t \le 1 - IDel_t \qquad t = 1, ..., |n|$$

$$Del_t \ge 0 \qquad t = 1, ..., |n|$$

$$Ret_t \ge 0 \qquad t = 1, ..., |n|$$

So that it can be expressed as a DPP, considering a partition of the *closing balance* of day t into five intervals that constitute the states of each *stage* because, although *cash balances* are bounded, they belong to a very dense set that is complex to consider without grouping. Thus, the intervals are defined as follows:

$$X_{t1} : \left[SS, SS + (MS - SS)/5\right)$$

$$X_{t2} : \left[SS + (MS - SS)/5, SS + (MS - SS) * (2/5)\right)$$

$$X_{t3} : \left[SS + (MS - SS)(2/5), SS + (MS - SS)(3/5)\right)$$

$$X_{t4} : \left[SS + (MS - SS)(3/5), SS + (MS - SS)(4/5)\right)$$

$$X_{t5} : \left[SS + (MS - SS)(4/5), MS\right]$$

Let us say that the bounds for each state X_{tk} are X_{tkd} and X_{tku}, respectively.

Alternatives are directly related to states, so they will also be defined as intervals. The costs associated with each *alternative-state* pair will be calculated from the midpoint of the interval of the *state* and the corresponding *alternative* point, except in cases where the transition from the previous *state* point to the current one can be given by an *alternative* equal to 0. In these cases, the associated cost represents that there is no *delivery* or *return*.

In order to simplify notation, the *alternative* k in stage t will be represented as Del_{tk}. If its value is positive, it refers to a *delivery*; if its value is negative, it refers to a *return*. Del_{tk} is defined by the difference between *state* s and the *closing balance* of day t–1, updated with the demand approach and with the *addition to ATMs*, for day t, i.e.:

$$Del_{tk} = X_{ts} - X_{t-1,s'} - Dep_t + Wth_t + AdiATM_t, \quad s = 1, 2, ..., 5 \quad s' = 1, 2, ..., 5$$

where X_{ts} and $X_{t-1,s'}$ represent the s^{th} and s'^{th} state of X, in stages t and t – 1, respectively. Thus:

- In *stage* 1:

  The *opening balance* is represented by X_0, therefore there are five possible *alternatives*:

$$Del_{11} : \left[X_{11d} - X_0 - Dep_1 + Wth_1 + AdiATM_1, X_{11u} - X_0 - Dep_1 + Wth_1 + ADiATM_1\right)$$

$$Del_{12} : \left[X_{12d} - X_0 - Dep_1 + Wth_1 + AdiATM_1, X_{12u} - X_0 - Dep_1 + Wth_1 + ADiATM_1\right)$$

$$Del_{13} : \left[X_{13d} - X_0 - Dep_1 + Wth_1 + AdiATM_1, X_{13u} - X_0 - Dep_1 + Wth_1 + ADiATM_1\right)$$

$$Del_{14} : \left[X_{14d} - X_0 - Dep_1 + Wth_1 + AdiATM_1, X_{14u} - X_0 - Dep_1 + Wth_1 + ADiATM_1\right)$$

$$Del_{15} : \left[X_{15d} - X_0 - Dep_1 + Wth_1 + AdiATM_1, X_{15u} - X_0 - Dep_1 + Wth_1 + ADiATM_1\right)$$

  The cost associated with *state* s, because of the selection of *alternative* k is:

$$C_1(k, s) = \left(X_0 + \widehat{Del}_{1k} + Dep_1 - Wth_1 - AdiATM_1\right) *i + KO + \widehat{Del}_{1k} * CO,$$

  where

  $\widehat{Del}_{1k}$ is a value in the interval Del_{1k}, considering that it is zero if the interval contains it.

$$KO = \begin{cases} KD & \text{if } \widehat{Del}_{1k} > 0 \\ KR & \text{if } \widehat{Del}_{1k} < 0 \\ 0 & \text{if } \widehat{Del}_{1k} = 0 \end{cases}$$

$$CO = \begin{cases} CD & \text{if } \widehat{Del}_{1k} > 0 \\ -CR & \text{if } \widehat{Del}_{1k} < 0 \\ 0 & \text{if } \widehat{Del}_{1k} = 0 \end{cases}$$

- In stage t (> 1):

  The *opening balance* of day t is represented by $X_{t-1,s'}$ and, for the generation of *alternatives* on day t, it is necessary to consider their possible values (s′ = 1, 2, …, 5). However, for *states* X^{th} and $X_{t,h+1}$ (h = 1, 2, 3 and 4), of the five possible *alternatives* for each one, four are coincident, since the alternatives are constructed from the same demand approach and the same *addition to ATMs*. The only thing that changes is the difference between stocks of day t – 1 and stocks of day t, which is a multiple of the length of each interval, let us say δ = (MS – SS) / 5. Thus, the alternatives in the stage t are 9:

$$\text{Del}_{t1}:\left[X_{t1d}-X_{t-1,5u}-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\right.$$
$$\left.X_{t1u}-X_{t-1,5d}-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$
$$=\left[-5\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\right.$$
$$\left.-3\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t2}:\left[-4\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ -2\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t3}:\left[-3\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ -1\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t4}:\left[-2\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ -\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t5}:\left[-1\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ \delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t6}:\left[-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ 2\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t7}:\left[\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ 3\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t8}:\left[2\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ 4\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

$$\text{Del}_{t9}:\left[3\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t,\ 5\delta-\text{Dep}_t+\text{Wth}_t+\text{AdiATM}_t\right)$$

- The daily cost associated with *state* s, because of the selection of *alternative* k is:

$$C_t(k,s)=\left(\hat{X}_{ts}\right)*i+KO+\widehat{\text{Del}}_{tk}*CO,$$

where

$\widehat{\text{Del}}_{tk}$ is a value in interval Del_{tk}, considering it to be zero if the transition from the previous *state* point to the current one can be given with no orders; and $\hat{X}_{ts}$ is the state associated with $\widehat{\text{Del}}_{tk}$.

$$KO=\begin{cases}KD & \text{if } \widehat{\text{Del}}_{tk}>0\\ KR & \text{if } \widehat{\text{Del}}_{tk}<0\\ 0 & \text{if } \widehat{\text{Del}}_{tk}=0\end{cases}$$

$$CO=\begin{cases}CD & \text{if } \widehat{\text{Del}}_{tk}>0\\ -CR & \text{if } \widehat{\text{Del}}_{tk}<0\\ 0 & \text{if } \widehat{\text{Del}}_{tk}=0\end{cases}$$

Finally, the forward recursion of the solution to the problem is represented by function $f_t(s)$, which is equal to:

$$\min_k\left\{C_t(k,s)+f_{t-1}(h:X_{t-1,h}=X_{ts}-\text{Delt}_k\right.$$
$$\left.-\text{Dep}_t+\text{Ret}_t+\text{AdiATM}_t)\right\},$$

considering that:

$$f_1(s)=\min_k\left\{C_1(k,s)\right\}.$$

The best *alternative* in *stage* t for *state* s is the one whose associated cost is $f_t(s)$.

Based on the above definitions, it is possible to determine an efficient sequence of orders to the central vaults in terms of transfer costs and immobilisation of money. It is important to consider that the parameters of the problem should have a 'high' accuracy or the solution must be treated with some caution.

2.3 Proposal for parameters definition

This section defines the parameters needed to solve the *cash management at bank branches* problem in the way we described in section 2.2.

2.3.1 Demand for cash transactions

Forecasting demand for cash transactions is necessary, not only to predict *deposits* and withdrawals in order to ensure the *service level*, but also to establish tactics, setting up the *safety stock* and *maximum cash storage* for branch vaults, and the *minimum* and *maximum amount of money in the ATMs*, that yields more objectivity in transfer orders, which leads to potential savings.

For this purpose, a fit of the proposed model is presented in this section, measuring the level of accuracy achieved in a considerable number of branches. To this end, we give a brief description of the require theory.

2.3.2 Compound process

The $\{T_m, Y_m\}$ pair is called a *compound process*. $\{T_m\}$ represents the sequence of time events in a stochastic process, $\{Y_m\}$ is the set of random variables associated with $\{T_m\}$, mutually independent; and N(0, t] is the corresponding counting process, i.e., m = 1, …, N(0, t].

The random sum of a *compound process* is defined as:

$$S(0,t]=\begin{cases}\sum_{m=1}^{N(0,t]}Y_m & N(0,t]\geq 1\\ 0 & N(0,t]=1.\end{cases}$$

Depending on the probability distribution of N(0, t], the distribution of the random sum and the *compound process* take their names. That is, if N(0, t] has a *negative binomial* distribution, the distribution of S(0, t] and the process are called: *compound negative binomial* and *compound negative binomial process*, respectively.

In the case of a compound Poisson process, the expected value of the random sum is:

$$E\left(S(0,t]\right)=\left(\int_0^t\lambda(s)ds\right)*\left[E\left(Y_m\right)^2+\text{Var}\left(Y_m\right)\right].$$

Variance of the random sum:

$$\mathrm{Var}\left(S(0,t]\right)=\left(\int_0^t \lambda(s)ds\right)*\left[E\left(Y_m\right)^2+\mathrm{Var}\left(Y_m\right)\right].$$

2.3.3 Generalised linear models

The *generalised linear models* (GLMs) are a family of models where response variable y can be quantitative or qualitative, assuming it has a distribution function that belongs to the exponential family, i.e., its density function can be expressed as:

$$f(y \mid \theta)=e^{(p(\theta)y-q(\theta)+g(y))},$$

where $p(\theta)$, $q(\theta)$, $g(y)$ are functions.

The components of the GLMs are:

- *Random component*: This is the response variable y. It is needed to define its probability distribution.
- *Systematic component*: This specifies the variables used in the linear predictor, which is the result of the linear combination of these variables and is selected for the construction of the model.
- *Link*: The link between the components defined above. It relates a monotonic function of the expected value of the response variable, $g(\mu)$, to the linear predictor. The simplest function $g(\mu)$ is the identity and it is called the identity link.

Some common GLMs are:

- *Poisson loglinear model*. This is a model for response variables in which values belong to the set of natural numbers. It assumes a Poisson distribution for the random component and uses the log function as a link.
- *Gamma loglinear model*. This is a model for response variables whose values belong to the set of non-negative numbers. It assumes a gamma distribution for the random component and uses the log function as a link.

Let μ be the expected value of variable y and x the explanatory variable, then the association between these variables is a *loglinear model*, as mentioned above, and has the following representation:

$$\log\mu=\alpha+\beta x$$
$$\Rightarrow \mu=e^{\alpha+\beta x}=e^{\alpha}\left(e^{\beta}\right)^{x}.$$

In addition to estimating β parameters, it is necessary to check the veracity of the assertions made about some unknown population characteristics. The procedure for this is known as *hypothesis testing*.

There are test statistics for the *significance* of variables, as well as others that test the accuracy with which the systematic component can describe the random component with the selected link, i.e., *goodness of fit*. Since the latter is a measure of the overall performance of the model, this is suggested as the main component when choosing one perspective or other. *Deviance* is a measure that helps to know about the adequacy of the model. Let L_M be the maximised log-likelihood value for the model in question and L_S the maximised log-likelihood value for the most complex model, i.e., the *saturated model*. The deviance of model M is defined as –2 times the logarithm of the likelihood ratio and in order to compare this model and the saturated one, we have:

$$\text{Deviance}=-2\left[L_M-L_S\right].$$

The purpose of the *deviance* is to test the hypothesis that all the parameters that are in model S but not in model M equal zero. For large samples, it has approximately a chi-square distribution with *degrees of freedom* equal to the number of parameters in model S but not in model M.

The *null deviance* is defined as the deviance when model M is just a constant.

We can compare two models: M_0 and M_A, none of them saturated but M_0 a special case of M_A, through their deviances:

$$-2\left[L_{M0}-L_{MA}\right]=M_0\text{Deviance}-M_A\text{Deviance}.$$

This test statistic is analogous to the *F-test* that compares linear regression models with normal distribution response variables.

The following ratio is called (Dobson, 2002) the *pseudo* R^2, since there is no a R^2 in GLMs:

$$(\text{Null deviance}-\text{Residual deviance})/\text{Null deviance}.$$

This ratio can be interpreted as the proportion of the variation in the response variable explained by the explanatory variables.

2.3.4 INLA

Integrated nested Laplace approximation (*INLA*) is a computational approach in the R software introduced by Rue et al. (2009). This approach performs Bayesian inference of the *latent Gaussian models* type, i.e., models whose density $p(x \mid \theta)$ is assumed Gaussian with mean equal to zero and a precision matrix $Q(\theta)$, where θ represents the vector of hyperparameters. In this way, distributions are as follows:

$$(\theta)\sim p(\theta)$$
$$(x \mid \theta)\sim N\left(0, Q(\theta)^{-1}\right)$$
$$\left(y_i \mid x, 0\right)\sim p\left(y_i \mid \eta_i, \theta\right),$$

where as we have already mentioned, θ are (hyper)parameters, $p(\theta)$ is typically taken to be non-informative, x is a latent Gaussian field, η is a linear predictor based on known covariate values c_{ij} $(\eta_i=\Sigma_j\ c_{ij^*}\ x_j)$, and y is a data vector. The joint distribution of the variables in the model is $p(y, x, \theta)$, that is a function of $(y_i \mid x, \theta)$, $Q(\theta)$ and $p(\theta)$. y is taken as fixed to get the posterior marginal densities of the latent variables $p(x_i \mid y, \theta)$ given a fixed hyperparameter value, then these marginals are integrated over the approximations of the hyperparameters' posterior density p(θ | y) (Cseke and Heskes, 2011).[4]

Thus, the INLA approach consists of: first, approximating the full posterior p(θ | y) (by using the Laplace approximation) that will be used later to integrate out the uncertainty with respect to θ when approximating the posterior marginal of x_i. The second step computes the Laplace approximation of the full conditionals $p(x_i \mid y, \theta)$ for selected values of θ. Finally, the approximation for the marginal of the latent variables is obtained $p(x_i \mid y)$.

Summing up, the main objective of the INLA approach is to get an approximation to the marginal posteriors for the latent variables as well as to the hyperparameters of the Gaussian latent model. It is important to mention that INLA uses accurate deterministic approximations instead of *Markov chain Monte Carlo* (MCMC) simulations in order to estimate the posterior marginal.

2.3.5 Description of the models

This starts from the premise that any bank has the following current and historical information:

- branch ID
- date of accounting record
- number (of transactions) of deposits
- amount (of money) of deposits
- number (of transactions) of withdrawals
- amount (of money) of withdrawals.

From this information, the following variables should be generated (the nomenclature is suggested as well):

- Branch: Branch ID (qualitative, nominal)
- Wkingday: Day of the week wherein the accounting movements were registered (qualitative, nominal) 1 = Monday, 2 = Tuesday,…, 5 = Friday.
- Mnthday: Day of the month wherein the accounting movements were registered (qualitative, nominal).
- Payday: Indicator variable for paydays, considering the majority of people. It is suggested to consider two consecutive working days (qualitative, nominal).
- Pstpayday: Indicator variable for working days after paydays. It is suggested to consider two consecutive days (qualitative, nominal).
- Holiday: Indicator variable for public holidays (qualitative, nominal).
- Pstholiday: Indicator variable for the working day after public holidays (qualitative, nominal).
- Month: Month wherein the accounting movements were registered (qualitative, nominal) 1 = January, 2 = February, …, 12 = December.
- Year: Year wherein the accounting movements were registered (qualitative, nominal).
- Deptxn: Number (of transactions) of deposits (quantitative, discrete).
- Depamnt: Amount of money from deposits (quantitative, continuous).
- Wthtxn: Number (of transactions) of withdrawals (quantitative, discrete).
- Wthamnt: Amount of money from withdrawals (quantitative, continuous).

A model is proposed for approaching demand based on the random sum of the *compound process* $\{T_{m,t}, Y_{m,t}\}$, that is associated with the counting process N(t–1, t], where:

- $T_{m,t}$ represents the time when the m^{th} transaction happens on day t.
- $Y_{m,t}$ represents the amount of money of the m^{th} transaction on day t.
- N(t–1, t] represents the number of transactions on day t.

It is important to note that we must distinguish transactions that represent *deposits* from those that represent *withdrawals*, considering that *withdrawals* are separated, as well as those carried out with tellers from those that occur in the ATMs.

Because of the characteristics of the data, it is natural to suggest, for the first component of the *compound process*, approaching the future data according to a *Poisson loglinear* model (Agresti, 1196), while the second component is based on a *gamma loglinear model* because of the flexibility of its probability distribution, which mips to represent a variety of distribution forms with only two parameters (Wilks, 1990).

The exploratory analysis is not presented because the proposed models are based on the experience of cash flow in bank branches and because these are models that generally represent counting variables and positive continuous variables, respectively. Proposed models, M1 and M2, only differ in the perspective of their approaches, i.e., both take the same random and systematic components, using the same link. M1 and M2 must be developed independently for each branch because of their possible specificity, e.g., a branch located in a highly commercial area behaves quite differently from a branch that is next to residential neighbourhoods: different amounts, trends, seasonality, etc. Models are simpler to treat this way, as it lowers the number of variables and the quantity of problems to be addressed (e.g., cross-correlation between agents at the same point of time (Gujarati and Porter, 2009), which should be reviewed in panel data).

Both models, M1 and M2 consist of approaching every component of the compound processes in detail, for *deposits* and *withdrawals*, separately, M1 through the use of GLMs while M2 is approximated through the use of the *INLA* function without specifying the prior distribution of the parameters.

Thus, regressions are performed according to the following statements, in the case of *deposits*.

Prior to the regression fit, indicator variables are generated for the (originally) nominal variables, considering that the number of indicators should be the possible categories in each nominal variable minus one. In an attempt to sum up the regression expressions, i.e., without all the indicator variables, here is the description for the number of transactions approach:

$$\begin{aligned} &\text{Deptxn} \sim \beta_0 + \beta_1 * \text{Wkingdayfact} + \beta_2 * \text{Mnthdayfact} \\ &+\beta_3 * \text{Payday} + \beta_4 * \text{Pstpayday} + \beta_5 * \text{Holidayfact} \\ &+\beta_6 * \text{Pstholiday} + \beta_7 * \text{Monthfact} + \beta_8 * \text{Yearfact}. \\ &(\textit{Poisson loglinear model}). \end{aligned}$$

Even though a frequent problem called *overdispersion* (variance > mean) can happen, we are not making any adjustment, as our objective is to approach with point estimates, and the *overdispersion* correction is not necessary.

The amount of money approach is:

$$\begin{aligned} &\text{Depamnt} \sim \beta'_0 + \beta'_1 * \text{Wkingdayfact} + \beta'_2 * \text{Mnthdayfact} \\ &+\beta'_3 * \text{Payday} + \beta'_4 * \text{Pstpayday} + \beta'_5 * \text{Holidayfact} \\ &+\beta'_6 * \text{Pstholiday} + \beta'_7 * \text{Monthfact} + \beta'_8 * \text{Yearfact}. \\ &(\textit{Gamm loglinear model}). \end{aligned}$$

Similarly, *Wthtxn* and *Wthamnt* are approximated by the definition of the values of the *compound process* for *withdrawals*.

The big difference between M1 and M2 is the assumption about the parameters. M1 only assumes randomness in the response variable, not in the regression parameters as M2 does.

2.3.6 Model selection

Below, the illustration of a model fit is explained in order to show how regressions (described above) could approach real data. Only the case of *number of deposits* in a particular branch is shown. Figure 2 shows the actual behaviour of deposits. Figure 3 illustrates the approach from a *frequentist* perspective, while in Figure 4 we took our approach from a *Bayesian* perspective.

The model was fitted to determine the components of the *compound process*, both for *deposits* and for *withdrawals*. It was fitted for 30 real branches, considering the data for a whole year, and evaluating the performance of both perspectives: *frequentist* and *Bayesian*.

In all the cases, approaching this problem through a conventional *GLM*, i.e., in a *frequentist* manner, we got a better performance of the model.

Trying to compare regressions fitted with INLA vs. regressions with traditional GLM, we realised that the *standardised residuals* were greater than those obtained using traditional GLMs. INLA fitted transactions reported a sum of squares of *standardised residuals* greater than 1,049. Traditional GLM did it with a sum of squares of *standardised residuals* since of 1.31.

Figure 2 Number of deposits for a branch in a year

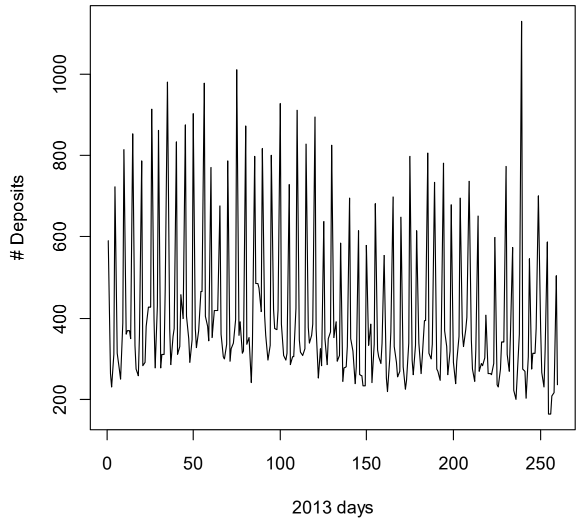


Figure 3 Number of deposits for a branch in a year (black) vs. traditional GLM approach (blue) (see online version for colours)

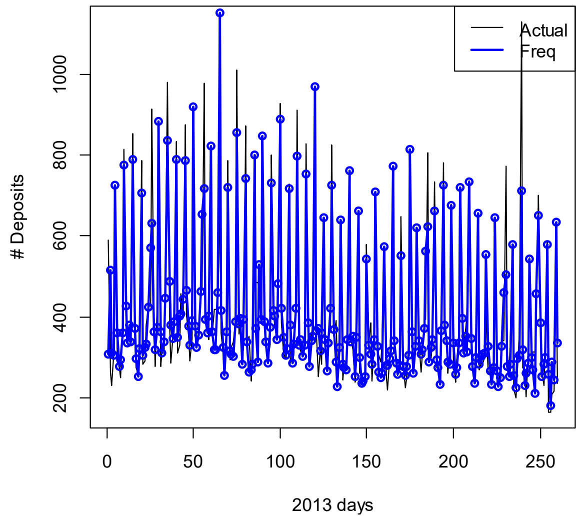
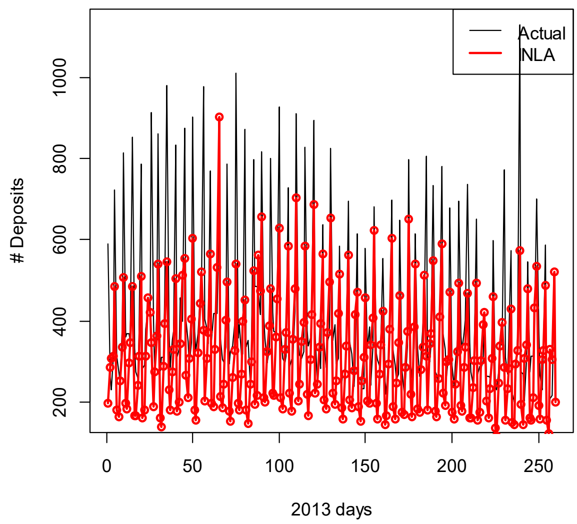


Figure 4 Number of deposits for a branch in a year (black) vs. INLA approach (red) (see online version for colours)



INLA fitted amounts reported a sum of squares of *standardised residuals* greater than 1,077. Traditional GLM did it with a sum of squares of *standardised residuals* since of 1.48.

To sum up, *model M1* performs better.

Once we detected as better option the GLM, we measure its *pseudo* R^2. It was up to 0.86, for transactions, and up to 0.82, for amounts; talking about deposits. Meanwhile for *withdrawals*, the *pseudo* R^2 was up to 0.86, for transactions, and up to 0.74, for amounts. Then, *model M1* is not just better model (vs. *model M2*) but it is also an acceptable model.

2.3.6.1 Safety stock and maximum cash storage

Once we have selected the classical *frequentist* GLM to approach *deposits* and *withdrawals*, we define the parameters that delimit the *closing balances*. The proposed way of defining them is based on the demand approach in addition to feedback from the cash administrator in the branch and feedback from the *control and risk officer*; however, feedbacks (in both cases) have their own and potentially varied criteria that will not be addressed in this work, so they will only be expressed, without further analysis.

A 'wide' range, resulting from the assignment of a *safety stock* and *maximum cash storage*, is critical for allowing an optimal cash flow through orders; in fact, ideally only the *safety stocks* could be set up to avoid cash unavailability and thus, *maximum cash storage* would be implicitly determined in the costs minimisation. But, it is necessary to set the maximum cash storage, mainly for security reasons and also to give a required input for our proposed solution.

Since these parameters are generally updated once a year, we propose that preliminary values be defined as stated below:

- *Safety stock*. If known, it is set as the necessary amount to maintain the service while it is applied a *delivery*. If unknown, it is proposed to take a percentage of $P\Pi_q$, which is defined as a proportion of the q percentile of the amount of daily *withdrawals* (Π), taken from the demand for cash transactions approach. The percentage to be considered should represent the fraction of the day that it takes to receive a *delivery* plus the time spent on counting and confirming this cash into the system. For example, if the working hours in a branch are from 9 AM to 4 PM, a delivery arrives at 10 AM and the cash administrator spend 30 minutes counting and confirming it into the system, the percentage to be considered will be (using the 24 hour clock): [(10 – 9) + 0.5] / (16 – 9), i.e., 21.42%). P is a factor that represents the fraction of *withdrawals* that are not solved by *deposits*, i.e., (*withdrawals* – *deposits*) / *withdrawals*, and it can be estimated using the average of daily factors, by applying the demand approach of the year to be configured. Owing to the criticality of cash (un)availability, in order to obtain P, we propose to only consider the daily factors with positive values and to define q with the highest possible value, i.e., q = 1.
- *Maximum cash storage*. It is easy to see that this parameter is less important than the safety stock parameter for guaranteeing cash availability, so we propose to fix it based on the capacity of the security equipment (vault, strongbox, etc.) to keep the cash safe, checking that it is enough to store the *safety stock* in addition to the estimated demand for every day of the year to be configured. It is important to consider that this capacity, expressed in an amount, is variable owing to possible denominations for bills and coins.

In both parameters, the final values must be given after the feedbacks, mentioned at the beginning of this section.

In an analogous manner, another two parameters are going to be defined: the minimum and maximum amount of money in the ATMs, complementing the demand approach with feedbacks from the devices vendors.

2.3.7 Other parameters

Apart from the above, there are other parameters whose value is yet to be defined. These are:

X_0 *opening balance* of the first day to be evaluated, i.e., closing balance of day 0

i daily cash immobilisation cost, including opportunity cost and risk fee

KD fixed cost per *delivery*

KR fixed cost per *return*

CD variable cost per *delivery*

CR variable cost per *return*.

As we pointed out earlier, it is important to make these parameters precise, by updating their values according to the latest information.

3 Vault policy

The description of the problem based on a *dynamic programming* perspective and the set up of the parameters as proposed in Section 2 constitutes the main input for the *vault policy*. The rest consists of a set of rules suggested by cash administrators in branches and, most importantly, follows intuition. The complete policy consists of:

1. Obtaining a *deposits* and *withdrawals* forecast for 'short' horizons. We suggest a monthly time horizon, because this is the minimum frequency used for expense evaluations (http://www.referenceforbusiness.com/encyclopedia/Bre-Cap/Budgeting.html).
2. When applicable, setting the annual parameters: *safety stock*, *maximum cash storage*, *minimum* and *maximum amount of money in the ATMs*.

3 Based on the *withdrawals* forecast for ATMs, establishing the $AdiATM_t$ plan. We suggest making additions to ATMs before high demand days.

4 Configuring the remaining parameters and solving the problem with a *dynamic programming* perspective, using the definition of states, stages, alternatives and recursion described in 2.2.

5 The set of orders that are obtained per day is the original plan of recommendations.

6 Recommendations are adjusted every day, based on the following rules:

Let:

Re be the recommendation obtained in the preceding step

B be the *closing balance* obtained by following the recommendation of the day and considering the demand forecast.

O be the final order, i.e., the adjusted recommendation.

If not explicitly indexed, every one of the above refers to its value on day t.

Rules:

```
Case Re ≥ 0:
  Case SS ≤ B ≤ MS:
    If B-Re ≥ SS, then
      O ← 0, else
      O ← Re
  Case B ≤ SS:
    O← Re + Average(SS, MS)-B
  Case B ≥ MS:
    If B-Re ≥ MS, then
      If Re_{t+1} > 0, then
        O ← 0, else
        O ← Re + Average(SS, MS)-B,
else
      If B-Re≤ SS, then
        O ← Re, else
        O ← 0
Case Re < 0:
    Case SS≤ B≤ MS:
      If B-Re ≤ MS, then
        O ← 0, else
        O ← Re
    Case B≤ SS:
      If B-Re≤ SS, then
        O ← Re + Average(SS, MS)-B, else
        If B-Re≥ MS, then
          If Re_{t+1} > 0, then
          O ← 0, else
            O ← Re + Average(SS, MS)-B,
  else
      O ← 0
    Case B ≥ MS:
      If Re_{t+1} > 0, then
        O ← 0, else
        O ← Re + Average(SS, MS)-B
```

These rules make it possible to react to inconvenient balance deviations caused by errors of estimation.

4 Results

Before discussing the main findings, it is important to note how we validated and verified the *vault policy*. This proposal was designed, programmed and reviewed by a team of *subject matter experts* (*SMEs*) that we belong to. In order to complete the validation, we made an exhaustive exploration of cases, in particular, of critical cases, i.e., those where the *maximum cash storage*, sufficiency of balances, cash transfer needs, were proven. In general, all explorations were done with the use of the *MILP* model, presented in Section 2.2. This model was useful, even knowing possible (bounded) differences, as it is a relaxation of the *dynamic programming* model that is done by not considering the last one's range definitions. Additionally, some controls were inserted to verify consistency during implementation. We certified that the detailed *closing balance* is contained in the range of the corresponding *state* (that includes the *maximum cash storage*), the orders were applied respecting the *alternatives* definition, and the *principle of optimality* of Bellman was fulfilled.

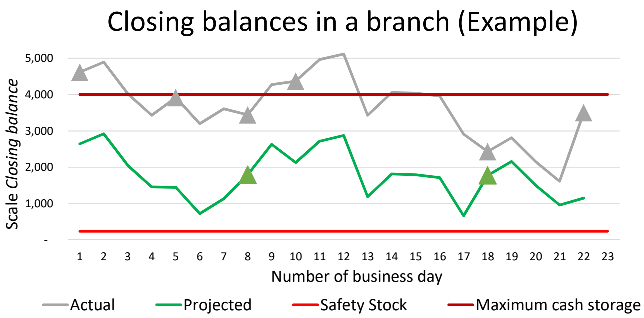
In a *desktop exercise*, the *vault policy* was applied to 30 branches, the same branches used for model selection in Section 2.3. It is important to point out that these branches have been 'optimised' by a very important supplier with worldwide presence.

We set all the parameters according to this proposal. The demand of cash transactions approach and all the parameters linked to it were constructed using the history of one and a half years. Then, the *vault policy* was applied to get information for a month, i.e., generating 690 observations, and assessing its performance in contrast with actual data.

The next graph gives an example of the behaviour of a vault.

Figure 5 shows the *closing balances* (in scale for reasons of confidentiality) of one of the 30 branches. There, we can see that the actual number of deliveries (grey triangles) was six in this month, whereas with the *vault policy* only two transfers (green triangles) would be required. It seems that on day ten it was absurd, even at the beginning of the day, to do a delivery. It is clear that in the current scheme there is an opportunity in respect of the *maximum cash storage*, probably because of the accuracy of the demand forecast. One important aspect of the *vault* *policy* is that you gain in efficiency when recommendations are complemented with the aforementioned rules.

Figure 5 Closing balances in a branch (grey) vs. projected with *vault policy* (green) (see online version for colours)



Looking at the bigger picture, the branch in the example also needs to review the definition of the *safety stock* in the current model, it does not need to keep a lot of money, and more if it comes from cash *deliveries*. It is incurring double and unnecessary costs.

The case shown in Figure 5 is a particular one; this means that every case can exhibit different improvements when the *vault policy* is applied. However, here are the main highlights from the *desktop exercise* for the 30 branches:

- Total shipments: 26.3% less than actual.
- Total transferred amounts: 0.6% less than actual.
- Cash storage: 29.8% above the actual.
- Balance under SS: equal to actual (6/690).
- Balance over MS: 2/690 less than actual (27/690).
- Total savings are estimated in 0.9%, based on a costs approach, also for actual transfers and closing balances.

This summary shows that, for this group of branches and considering representative costs for the Mexican market, this proposal tends to store more cash than the current scheme besides requesting more money per transfer. For this instance, we obtained savings that can represent serious money for many banks while respecting constraints, at least as much as current schemes.

Values for both immobilisation and transfer costs have an imminent impact on the results. For this 30-branches case and with reference to the current scheme, we achieved savings by lowering the number of transfers. Surely, we would not have got the same results with other cost values or with another 'current scheme'.

5 Conclusions

This paper is a proposal that makes it possible to manage bank branches, in terms of cash, with realistic assumptions that are easy to implement. Although this topic has been written about, the *vault policy* being presented has performed well, as confirmed by the evaluation of 30 branches, optimised by a global tech company.

The *vault policy* describes tactics for approaching cash optimisation in bank branches. One component of this policy is based on statistical models: *GLMs* through frequentist and Bayesian methods, as well as a *dynamic programming* model. However, the other important component is human experience. This complement gives us positive results.

The statistical models gave the elements of the *compound process*. It was defined in order to approach the demand of cash transactions in the horizon, where we got better performance from a *frequentist* perspective rather than a *Bayesian* perspective, using INLA.

INLA is a recent command in the R software. It uses accurate deterministic approximations instead of *MCMC* simulations to estimate posterior marginals.

Based on the demand for cash transactions, we can set many parameters needed to implement the *vault policy* (in Section 2, we suggest how to update them all). Once all the parameters have been configured, we can run the first recommendations, which are going to be adjusted on a daily basis by a set of rules, defined with the help of cash administrators, in order to react to balance deviations resulting from the forecast errors. As we have already mentioned, the performance of this *vault policy* was measured for 30 branches using some indicators, presented in Section 4. The results obtained reflect the impact on an instance, i.e., the implementation of the *vault policy* does not imply a certain behaviour of the cash balances and transfers. Depending on the cost values, that are valid for a specific case, the results may show a tendency to immobilisation or non-immobilisation, to do more/fewer transfers with a greater/smaller amount of money per transfer.

For future developments, we suggest including possible branch networks in order to setup routes that represent greater savings in the entire network than those obtained in an individual perspective. This would be even further improved if there could be cash exchanges between the branches, with no intermediation of central vaults. For this, it is necessary to understand how to establish the validation processes directly at branches.

Acknowledgements

In special appreciation to *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica*, PAPIIT, which has enabled the development of this work.

References

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*, pp.71–97, John Wiley & Sons, Inc., New York.

Bell, P.C. and Hamidi-Noori, A. (1984) 'Foreign currency inventory management in a branch bank', *The Journal of the Operational Research Society*, Vol. 35, No. 6, pp.513–525.

Cabello, J.G. (2013) 'Cash efficiency for bank branches', *SpringerPlus*, Vol. 2, p.334, ISSN: 2193-1801.

Chen, M-H. (1991) *How Much Cash Should a Bank Maintain?: An Optimization Approach*, MS thesis, Massachusetts Institute of Technology, Sloan School of Management, ID 26147907.

Cseke, B. and Heskes, T. (2011) 'Approximate marginals in latent Gaussian model', *Journal of Machine Learning Research*, Vol. 12, pp.1–39, ISSN: 1532-4435.

Dobson, A.J. (2002) *An Introduction to Generalized Linear Models*, 2nd ed., pp.43–82, 151–168, Chapman & Hall, Boca Raton.

Girgis, N.M. (1968) 'Optimal cash balance levels', *Management Science Theory Series*, Vol. 15, No. 3, pp.130–140.

Gujarati, D.N. and Porter, D.C. (2009) *Econometría*, 5th ed., pp.591–616, McGraw Hill.

http://www.referenceforbusiness.com/encyclopedia/Bre-Cap/Budgeting.html (accessed 7 January 2015).

Lin, Y. (2002) *General Systems Theory. A Mathematical Approach*, pp.135–161, Springer, USA, Series ISSN 1574-0463.

Rue, H., Martino, S. and Chopin, N. (2009) 'Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations', *JRSS-series B (with discussion)*, Vol. 71, No. 2, pp.319–392.

Simutis, R., Dilijonas, D., Bastina, L., Friman, J. and Drobinov, P. (2007) 'Optimization of cash management for ATM network', *Information Technology and Control*, Vol. 36, No. 1A, pp.117–121, ISSN: 1392-124X.

Wilks, D. (1990) 'Maximum likelihood estimation for the gamma distribution using data containing zeros', *Journal of Climate*, Vol. 3, pp.1495–1501, ISSN: 0894-8775,

Notes

1 Note that now the objective function does not include Dep_t, Wth_t or $AdiATM_t$ since they are considered constants that could be simplified.