APPRAOCHING DEMAND OF CASH TRANSACTIONS AT BANK BRANCHES

Miguel Aguilar Zaragoza\(^{(a)}\), Idalia Flores de la Mota\(^{(b)}\)

\(^{(a)}\)National Autonomous University of Mexico  
\(^{(b)}\)National Autonomous University of Mexico

\(\text{maguilarz@live.com, idalia@unam.mx}\)

ABSTRACT
Trying to maintain a service level in many operations can result very expensive, especially if there are severe penalties due to a low level related with the service level objective. In such cases, we seek to achieve a goal even at the cost of significant loss in the operation efficiency. In particular, bank companies at their branches undertake transactions related to cash money as main service, so, not having available cash becomes a critical issue, which slightest consequence is the bank company prestige deterioration, and not only affectations for a specific branch.

This paper has the purpose of establish an approach for cash transactions demand at bank branches, with perspectives that best fit, which helps to control the risk of stocking out of cash, setting up some parameters, \(e.g.\) branches’ safety stock of cash, and projecting horizons of cash balances, given certain scenarios.

The model we present is developed with frequentist and Bayesian perspectives, the last one is a ‘recent’ develop that covers in a more efficient way, the Markov Chain Monte Carlo (MCMC) simulations. So, by means a real evaluation, it will be estimated the accuracy level that can be got approaching the cash demand with each of these, in particular with the best.

Keywords: bank branches, Compound Process, INLA, Generalized Linear Models, simulation.

1. INTRODUCTION
Within their administration, banks have the ongoing task of developing their activities seeking to be efficient, \(i.e.\), fulfilling the service (level) at minimal cost. Among the main activities, banks offer services through their teller desks and automated devices that generate cash inflows and outflows in their branches.

The purpose of finding the efficient cash management in bank branches means to minimize money-transfer cost, opportunity cost (caused by the money storage in the vaults of the branches), and the amount exposed to risk of theft, maintaining the operation of this entities with a desired level. All these decisions are better supported with a truthful approach of cash transactions. Since it is the most representative proportion of cash services, and therefore “more predictable”, the scope of this article is only local currency, so it does not include foreign currencies.

Thus, the objective is to propose a model that helps us to know about the amount of money we need in every branch in order to facilitate, to whom are in charge of the branch vault, to manage it in an appropriate manner through an efficient handling of orders required to central vaults.

It has been written about the efficient cash management in bank branches, such as the paper of Julia García (2013), however, it has the lack of not having completely realistic assumptions, and it could be obtained solutions with unexpected consequences, including negative stop costs (transfer costs). Also, some proposals have been published for the management of foreign currencies in bank branches. In his paper, Bell (1984) discusses this issue as one of inventory management and results are shown through the use of decision rules, for any branch with a considerable volume of transactions with foreign currencies. Nevertheless, the foreign exchange demand is very singular; furthermore, as it is mentioned, it is not easy to implement these rules. Nonetheless, it is attractive the treatment of the vault stocks as an inventory. Also, proposals have been arisen to determine the optimal stock levels that ensure available cash to cover future expenses of any entity. Nadia Makary (1968) proposes an optimal inventory policy for the case where costs of money storage and lack of it, are convex functions, and in the scenario when the decision to increase or decrease inventory levels is with no fixed cost. All these efforts have addressed their problems giving little attention to model demand with a substantial accuracy. Since we suggest that some parameters and scenarios, for branches, should be set up based on the cash transactions...
demand, we need to approach it with more accuracy.

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

2. DESCRIPTION OF THE PROBLEM

The cash management at bank branches problem is addressed considering the situations contained in this section.

2.1. Context

Branch administrators must handle the cash flow of the vaults, according to the following diagram:

![Cash Flow Diagram](image)

Fig. 1: Cash inflows/ outflows in vaults of bank branches

In Fig. 1 possible movements in the cash stock in a branch are represented. 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 teller desks; Deliveries and Emergency deliveries are cash supplies from central vaults to the branch, with different sense of urgency; Returns are shipments 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 the stock at the end of the day, obtained by doing arithmetic operations with the prior concepts.

2.2. Variables and parameters

Traditional mechanisms of control for handling cash in banks, with branches that have a high level of discretionary to take decisions related to ordering deliveries and returns to central vaults, are based on establish bounds to the maximum number of transfers and the minimum amount per transfer, both for cost purposes; and to the maximum cash storage, for security purposes. Assuming that the first two bounds allow an efficient operation, it is easy to see that the third one owns the power to affect the good approach of the other two parameters. Also, it is easy to notice that it could be inconsistency between the cost-purposes and security-purposes bounds. That’s why it is important that maximum number of transfers and minimum amount per transfer are set once established the maximum cash storage. Even in our case (that avoids the cost-control parameters), it is very important to define a suitable maximum level to store cash, that allows efficiency while respecting the security parameter.

Besides the maximum cash storage, it is suggested to maintain a safety stock, both considering the available information. Consequently, in this paper it is proposed to define this interval (of amounts of money allowed for a branch) based on historical data and the feedback of the branch administrator to determine the lower bound, and the historical data and the feedback of the Risk/Control Office to determine an upper bound. In an analogous manner, other two parameters are going to be defined: the minimum and maximum amount of money in the ATMs, complementing the historical data with suggestions of the devices vendors.

Of course, deposits and withdrawals are unknown, but future demand is being estimated with a forecast model grounded in a Compound Process. So, the unique variables considered for this model are: number of transactions and transactions amount of money, both for deposits and withdrawals.

2.3. Problem objective

The main idea to solve the problem discussed is to generate an approach for cash transactions demand that allows, 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 the event and amount of cash in transfer orders, with a huge potential for cost savings.

It is going to be presented the best perspective of the model proposed, measuring the accuracy level reached in a considerable number of branches.

3. APPROACHING CASH TRANSACTIONS DEMAND

The demand approach, that is the base of our proposal, is being taken from the perspective that better fits of the following: Model M1 and Model M2, described below. Before that, we show a brief of the needed theory.

3.1. Compound Process

It is called Compound Process to the pair \{N(0, t], Y_i\}, where \(Y_i\) is the random variable associated with the \(i\)th occurrence of the counting process. These two variables are independent each other. It is defined the random sum of a Compound Process as:

\[
S(0, t) = \begin{cases} 
\sum_{i=1}^{N(0, t]} Y_i & \forall \ N(0, t] \geq 1 \\
0 & \ N(0, t] = 0 
\end{cases}
\]

Depending on the probability distribution of \(N(0, t]\), the distribution of the random sum and the
Compund 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:

\[ E(S(0, t]) = \int_0^t \lambda(s) \, ds \cdot E(Y). \]

Variance of the random sum:

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

3.2 Generalized Linear Models

Generalized Linear Models (GLMs) are a family of models which response variable, \( y \), may be quantitative or qualitative, assuming it has a distribution function that belongs to the exponential family, i.e., that its density function can expressed as:

\[ f(y|\theta) = e^{p(\theta) - q(\theta) + g(y)}, \]

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

The components of the GLMs are:
- **Random component**: Is the response variable \( y \). We need to define the probability distribution that has the same.
- **Systematic Component**: It specifies the variables used in the linear predictor, result of the linear combination of 1 and the explanatory variables selected for the model construction. Within these explanatory variables it could be the interaction of them.
- **Link**: The linkage between the components defined above. It relates a monotonic function of the expected value of the response variable, \( g(\mu) \), with the linear predictor. The simplest function \( g(\mu) \) is the identity and it is called identity link.

3.2.1 Poisson Loglinear Model

It is a model for response variables which values belong to the set of natural numbers. It assumes a Poisson distribution for the random component and it uses, as link, the log function.

Let \( \mu \) be the expected value of the Poisson variable \( y \) and \( x \) the explanatory variable, then the association between these variables in a Poisson Loglinear Model has the following representation:

\[ \log \mu = \alpha + \beta x \]

\[ \mu = e^{\alpha + \beta x} = e^\alpha (e^\beta)^x. \]

3.2.2 Gamma Model

It is a model for response variables which values belong to the set of non-negative numbers. It assumes a Gamma distribution for the random component and it uses, as link, the inverse function.

3.2.3 Hypothesis testing for GLMs

In addition to estimating \( \beta \) parameters, it is necessary to review the veracity of the assertions that are made regarding to some unknown population characteristics. The procedure to do this is known as hypothesis testing.

There are test statistics for significance of variables, as well as others that test the accuracy in which the systematic component can describe the random component with the selected link, i.e., goodness of fit. Since the last one is a measure of the overall performance of the model, it is suggested to be the criterion when choosing one perspective or other.

3.2.3.1 Deviance

The deviance is a measure that summarizes the model adequacy. Let \( L_0 \) be the maximized log-likelihood value for the model of interest and \( L_s \) the maximized log-likelihood value for the most complex model, i.e., the saturated model. The deviance of a model \( M \) is defined as \(-2\) times the logarithm of the likelihood ratio to compare the \( M \) model and the saturated:

\[ \text{Deviance} = -2[L_M - L_s]. \]

The deviance has the purpose to test the hypothesis that all parameters that are in model \( S \) but not in model \( M \) equals 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 the model \( M \).

The Null deviance is defined as the deviance when Model \( M \) is just a constant.

To compare two models: \( M_0 \) and \( M_A \), none of them saturated but \( M_0 \) a special case of \( M_A \), we can do it through their deviances:

\[-2(L_{M0} - L_{MA}) = M_0 \text{Deviance} - M_A \text{Deviance}. \]

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

The following ratio is called\(^1\) 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.

3.3 INLA\(^2\)

INLA (Integrated Nested Laplace Approximation) is a computational approach in the \( R \) software introduced by Rue et al. (2009). This approach performs Bayesian inference in the class of Latent Gaussian Models, i.e., models which density \( p(x|\theta) \) is assumed Gaussian with mean equals zero and with precision matrix \( Q(\theta) \), where \( \theta \) represents the vector of hyperparameters. Thus, distributions are in the following form:

\[ (\theta) \sim p(\theta) \]

\[ (x|\theta) \sim N(0, Q(\theta)^{-1}) \]

\(^1\) Dobson (2002)

\(^2\) Lingren, Finn., Rue, Havard, 2014. Bayesian Spatial Modelling with R- INLA
(y|x, θ) = p(y|θ),

where, as we mentioned, θ are (hyper)parameters, p(θ) is typically taken to be non-informative, x is a latent Gaussian field, η is a linear predictor based on known covariate values c_j (η_i =Σ_j c_{ij} x_j), and y is a data vector. The joint distribution of the variables in the model is p(y|x, θ), that is function of (y|x, θ), Q(θ) and p(θ). It takes y fixed to get the posterior marginal densities of the latent variables p(x|y, θ), given a fixed hyperparameter value, then it is integrated these marginals over the approximations of the hyperparameters posterior density p(θ|y). \(^3\)

The principal 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.

The INLA approach consists in approximate 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|y, θ) for selected values of θ. Finally, the approximation for the whole variables is obtained p(x|y).

Summing up, INLA uses accurate deterministic approximations instead of Markov Chain Monte Carlo (MCMC) simulations in order to estimate posterior marginals.

### 3.4 Models description

It 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 the information, mentioned above, the following variables should be generated (the nomenclature is suggested as well):

- Branch: Branch ID (Qualitative, nominal)
- Wknday: Day of the week wherein the accounting movements were registered (Qualitative, nominal)
  - I= 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 1 or 2 consecutive days (Qualitative, nominal)

- Pstpayday: Indicator variable for working days after paydays. It is suggested to consider 1 or 2 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)
- I= January, 2= February,…, 12= December
- Deptxn: number (of transactions) of deposits (Quantitative, discrete)
- Depamnt: amount of money from deposits (Quantitative, continuous)
- Wthamnt: amount of money from withdrawals (Quantitative, discrete)
- Wthamt: amount of money from withdrawals (Quantitative, continuous)

The paradigm for both models, M1 and M2, is to define, for each branch, the random sum of a Compound Process \{N(t_i, t_{i+1}), Y_{i,i+1}\}, where:

- N(t_i, t_{i+1}) represents the number of transactions between t_i and t_{i+1},
- Y_{i,i+1} represents the amount of money of the N(t_i, t_{i+1}) transactions performed.

At first, it is understood that the difference between t_i and t_{i+1} is one day, i.e., N(t_i, t_{i+1}) represents the number of transactions on day i, and Y_{i,i+1} represents the average amount of the transactions performed on day i.

Importantly, we must distinguish transactions that represent withdrawals from which represent deposits. Thus, it will be a model to estimate the number of deposits: \{N_d(t_i, t_{i+1}), Y_{d,i,i+1}\}, and one for withdrawals: \{N_w(t_i, t_{i+1}), Y_{w,i,i+1}\}.

Models, M1 and M2, must be developed in an independent way, talking about branches, because of the particularity that could exist in each one, e.g., a branch placed in a zone of high commercial activity, has a distinct behavior in relation to a branch located next to residential neighborhoods: different amounts, trends, seasonality, etc. In this way, models are simpler to treat, since it decreases the number of variables and the quantity of problems to address (e.g., cross-correlation between agents at the same point of time\(^4\), which should be reviewed in panel data).

#### 3.4.1 Model M1

M1 consists in approaching, punctually, every component of the processes: \{N_d(t_i, t_{i+1}), Y_{d,i,i+1}\} and \{N_w(t_i, t_{i+1}), Y_{w,i,i+1}\}, through the use of GLMs. Because of the features of data, it is natural to suggest, for the first component, to approach the future data according to a Poisson Loglinear

---


Model\(^5\), while the second component based on a Gamma Model, because of its probability distribution flexibility that allows to represent a variety of distribution forms with only two parameters\(^6\).

Thus, regressions are performed according to the following statements, for the deposits case:

Previous to the regression fit, there are generated indicator variables for the (originally) nominal variables, considering that the number of indicators should be the possible categories in each nominal variable minus one. Trying to sum up the regression expressions, this is, without all the indicator variables, here is presented the description for the number of transactions approach:

\[
\text{Deptxn} - \beta_0 + \beta_1 \times \text{Wkngdayfact} + \beta_2 \times \text{Mnthdayfact} + \\
\beta_3 \times \text{Payday} + \beta_4 \times \text{Pstpayday} + \beta_5 \times \text{Holidayfact} + \beta_6 \times \text{Ps\_holiday} + \beta_7 \times \text{Monthfact}. \quad (\text{Poisson Loglinear Model})
\]

Although it could happen a frequent problem named overdispersion (variance>mean), we are not going to do any adjust, since our objective is to approach punctually, and the overdispersion correction doesn’t affect it.

The amount of money approach is:

\[
\text{Depamnt} - \beta'_0 + \beta'_1 \times \text{Wkngdayfact} + \beta'_2 \times \text{Mnthdayfact} + \\
\beta'_3 \times \text{Payday} + \beta'_4 \times \text{Pstpayday} + \beta'_5 \times \text{Holidayfact} + \beta'_6 \times \text{Ps\_holiday} + \beta'_7 \times \text{Monthfact}. \quad (\text{Gamma Model})
\]

In an analogous manner, they are approached \textit{Wthtxn} and \textit{Wthamnt} in order to determinate the values of the Compound Process for withdrawals.

3.4.2 Model M2

Bearing in mind randomness in the regression parameters, \textit{M2} approaches punctually each component of the processes: \{N\_d(t, t\_i+1), Y\_d,i+1\} and \{N\_s(t, t\_i+1), Y\_s,i+1\}, with a Bayesian perspective. It is used the INLA function (mentioned in 3.3 section) without a specification for the prior distribution of the parameters, considering the following fits\(^5\):

\[
\text{Deptxn} - \beta_0 + \beta_1 \times \text{Wkngdayfact} + \beta_2 \times \text{Mnthdayfact} + \\
\beta_3 \times \text{Payday} + \beta_4 \times \text{Pstpayday} + \beta_5 \times \text{Holidayfact} + \beta_6 \times \text{Ps\_holiday} + \beta_7 \times \text{Monthfact}. \quad (\text{Poisson Loglinear Model})
\]

\[
\text{Depamnt} - \beta'_0 + \beta'_1 \times \text{Wkngdayfact} + \beta'_2 \times \text{Mnthdayfact} + \\
\beta'_3 \times \text{Payday} + \beta'_4 \times \text{Pstpayday} + \beta'_5 \times \text{Holidayfact} + \beta'_6 \times \text{Ps\_holiday} + \beta'_7 \times \text{Monthfact}. \quad (\text{Gamma Model})
\]

In the same way, it is approached \textit{Wthtxn} and \textit{Wthamnt} in order to determinate the values of the Compound Process for withdrawals.

4. RESULTS

Below, it is exposed the illustration of a model fit in order to show how regressions (described above) could approach real data. It is shown only the case of number of deposits in a determined branch. In Fig. 2 there is the actual behavior of deposits, in Fig. 3 it is illustrated the approach with frequentist perspective, while in Fig. 4 we drew the approach with Bayesian perspective.

---


\(^{7}\) Regressions are described summing up parameters in case of multiplication with indicators variables.
It was fitted the model to determine the components of the Compound Process, both for deposits and for withdrawals. It was fitted the model for 30 real branches, considering 2013 data, and evaluating the performance of both perspectives: frequentist and Bayesian.

In all the cases, approaching through a conventional Generalized Linear Model, i.e., in a frequentist manner, we got a better performance of the model. The pseudo $R^2$ was up to .86, for $N_d(t_i, t_{i+1})$ regression, and up to .82, for $Y_{d,i,i+1}$ one.

With the same position for withdrawals, the pseudo $R^2$ was up to .86, for $N_w(t_i, t_{i+1})$ regression, and up to .74, for $Y_{w,i,i+1}$ one.

Trying to compare regressions fitted with INLA Vs regressions with traditional GLM, we realized that, for the 30 branches, standardized residuals were greater than the obtained with traditional GLMs. INLA fitted $Y_{d,i,i+1}$ with a sum of squares of standardized residuals greater than 1049. Traditional GLM did it with a sum of squares of standardized residuals since of 1.31.

INLA fitted $Y_{w,i,i+1}$ with a sum of squares of standardized residuals greater than 1077. Traditional GLM did it with a sum of squares of standardized residuals since of 1.48.

5. CONCLUSIONS
The purpose of this paper is to suggest a model approach for cash transactions demand at bank branches, with the intention to help control the risk of stocking out of cash and being efficient managing the cash.

The model proposed to approach the cash flow in a branch is to establish a Compound Process $\{N(t_i, t_{i+1}), Y_{i,i+1}\}$, one for deposits and another for withdrawals. Each of these components should be estimated by a regression model that explains both variables with others that represent days in the weak, days in the month, paydays, holidays and months, some of them with a lag, described in section 3.

Because of the shown evidence, we suggest to do regressions in a traditional way, we mean, using GLMs in order to approach $N$ and $Y$. In many areas it is used a cutting edge way to approach variables, we refer to Bayesian perspective using INLA, but to fit the model proposed it is not suggested.

We did not have access to real data for ATMs, but because of the nature of its operation, it is natural to suppose that it performs better, fitting withdrawals, with a frequentist perspective than Bayesian, as it happened for withdrawals in branches.

Approaches obtained with the frequentist perspective can be used to set the maximum cash storage and safety stock. Since both parameters are normally fixed once a year, we proposed to determine these as:

- For safety stock: Let $P$ be a percentile (defined by the central office of the bank according to the desired service level) of the daily difference between withdrawals and deposits, approached for a year with the Compound Process, and let $Q$ be a percentile (estimated with a percentage of withdrawals satisfied at the beginning of a day) of the daily withdrawals approached for a year, as well. Then, the safety stock proposed is $\max\{P, Q\}$.

  It is important to point out that if the branch administrator knows about an extraordinary withdrawal, it must be covered the extraordinary demand besides the safety stock.

- For maximum cash storage, let $R$ be the maximum of the daily deposits approached for a year, and $T$ the amount of money that Risk/Control Office approve for the branch. Then, the maximum cash storage proposed is $\min\{R,T\}$.

- Minimum amount of money in an ATM: let it be a percentile (estimated with a percentage of withdrawals satisfied in a day, previous to the replenishment) of the daily withdrawals approached for a year.

- Maximum amount of money in an ATM: let $P_{\text{ATM}}$ be a percentile (defined by the central office of the bank with the desired service level) of the daily withdrawals approached for a year and let $V$ be the maximum amount of money that vendors recommend to hold in a ATM of a specific model. Then, the maximum amount of money in an ATM will be $\min\{P_{\text{ATM}}, V\}$.
ACKNOWLEDGMENTS
This research was supported by UNAM PAPIIT grant IN116012.

REFERENCES
Lingren, Finn., Rue, Havard, 2014. Bayesian Spatial Modelling with R- INLA

AUTHORS BIOGRAPHY
*Miguel* holds a BCs degree in Actuary from ITAM, and is currently studying a Masters in Operations Research, at UNAM. At work, he has served to the public and private sector in many activities related to the optimization/continuous improvement in processes that have to do with product sales; passenger arrivals coming from foreign flights, at Mexican airports; and cash management in bank branches. Nowadays, he is making an effort to implement the content of this proposal in the global firm he is working for.

Statistic inference and stochastic processes are the main topics he is passionately interested in.

*Idalia* She received her Ph.D. in Operations Research at the Faculty of Engineering of the UNAM. She graduated Master with honors and received the Gabino Barreda Medal for the best average of her generation. She has been a referee and a member of various Academic Committees at CONACYT. She has been a referee for journals such as Journal of Applied Research and Technology, the Center of Applied Sciences and Technological Development, UNAM and the Transactions of the Society for Modeling and Simulation International. Her research interests are in simulation and optimization of production and service systems. She is a full time professor at the Postgraduate Program at UNAM.