Accepted Manuscript

Modelling Repayment Patterns in the Collections Process for Unsecured Consumer Debt: A Case Study

Lyn C. Thomas , Anna Matuszyk , Mee Chi So , Christophe Mues , Angela Moore

PII: S0377-2217(15)00837-1
DOI: 10.1016/j.ejor.2015.09.013
Reference: EOR 13232

To appear in: *European Journal of Operational Research*

Received date: 14 January 2014
Revised date: 31 July 2015
Accepted date: 8 September 2015

Please cite this article as: Lyn C. Thomas , Anna Matuszyk , Mee Chi So , Christophe Mues , Angela Moore , Modelling Repayment Patterns in the Collections Process for Unsecured Consumer Debt: A Case Study, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.09.013

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- We use Markov chain models of payment patterns to estimate recovery rates.
- Models allow optimisation of write off policies.
- Models tested using large portfolio of UK retail loans during a 12 year period.
- Results aid the management of collections particularly the write-off decision.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Modelling Repayment Patterns in the Collections Process for Unsecured Consumer Debt: A Case Study

Lyn C. Thomas[a*], Anna Matuszyk[b], Mee Chi So[a], Christophe Mues[a], Angela Moore[a]

[a]*Southampton Business School, University of Southampton, Southampton SO17 1BJ, United Kingdom.*

[b]*Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw, Poland.*

Abstract

One approach to modelling Loss Given Default (LGD), the percentage of the defaulted amount of a loan that a lender will eventually lose is to model the collections process. This is particularly relevant for unsecured consumer loans where LGD depends both on a defaulter's ability and willingness to repay and the lender's collection strategy. When repaying such defaulted loans, defaulters tend to oscillate between repayment sequences where the borrower is repaying every period and non-repayment sequences where the borrower is not repaying in any period. This paper develops two models – one a Markov chain approach and the other a hazard rate approach to model such payment patterns of debtors. It also looks at simplifications of the models where one assumes that after a few repayment and non-repayment sequences the parameters of the model are fixed for the remaining payment and non-payment sequences. One advantage of these approaches is that they show the impact of different write-off strategies. The models are applied to a real case study and the LGD for that portfolio is calculated under different write-off strategies and compared with the actual LGD results.

Key words: OR in banking; payment patterns; collection process; Markov chain models; survival analysis models

[*] Corresponding author.
E-mail addresses: l.thomas@soton.ac.uk (L. C. Thomas)
Telephone: +44 (0)23 8059 7718
Fax: +44 (0)23 8059 3844

ACCEPTED MANUSCRIPT

1. Introduction

There are two major reasons to model the collections process for the recovery of defaulted consumer debt. Firstly the regulations, incorporated in Basel II (BCBS, 2004) and Basel III (BCBS, 2011), on the risk capital that banks must hold required banks to estimate Loss Given Default (LGD) for each segment of their loan portfolio. LGD is the percentage of the debt at default that is still not collected at the end of the collection process. Basel Accord II (BCBS 2004) suggests three ways of modelling LGD: historical average, regression approaches and modelling the recovery process. For consumer debt, the historic average does not make much sense and the regression approaches lead to poor results with models in the literature having R-squared between 0.05 and 0.22. One reason for these poor results is the non-normal form of the LGD distributions but another significant reason is that LGD depends partially on the debtor's capacity and willingness to repay but also on the collection strategy. The models in this paper allow incorporation of the lender/collector's write-off strategy, which materially affects the resultant LGD. They also allow lenders to identify which among a set of write-off strategies will be most profitable over the whole debt portfolio. This is a second reason for modelling the collections process since lowering LGD affects who should get credit in the first place and at what price.

Default is defined as borrowers being 90 days overdue or there is evidence to the lender that the borrowers will not repay. Default triggers the collections process as the lender seeks to recover the debt. Most collections processes measure their success by the Recovery Rate (RR) they achieve, where *RR=1-LGD*.

The recovery rate depends not only on the debtors' capacity and willingness to repay but also on the lenders' actions and their collection policy. Previous models have ignored the lenders' influence in their models. One such collection action is to write off the loan and make no further attempt to collect. Writing-off is determined by the collector's expectation of future recoveries and the effort in collecting them. Such trade-offs can be used to determine whether the future expected recovery amount including recovery costs would be positive. Currently collectors make such write-off decisions subjectively and are often swayed by end of the quarter financial objectives or the pressure on the collections process. There is currently little modelling support for such actions. As well as estimating Recovery Rate (RR), the models presented here support collectors in assessing this trade-off between recoveries and the effort involved. This trade-off is influenced by the way debtors have already been repaying their debts, the cost of the collection process, and the likely duration until the debt will be repaid. The models allow collectors to have some data driven indication of which write-off policies are most profitable.

Provided the debtor is contactable, collections start with an agreement for the debtor to repay a fixed amount every period or to pay off the debt in one payment. What subsequently occurs is that there is an initial sequence of periods of non-payment while the agreement is put into place, followed by a sequence of periods of payment. This may stop and then a sequence of non-payment periods occurs until repayment restarts again. This can be repeated several times throughout the collection process as some of the real data examples in Figure 1 show. Alternatively the debt may be "cured" in that the repayments made cover the defaulted amount. In this paper we take "cured" to mean the debt is fully repaid, but a minor adjustment of the models would allow "cured" to mean a satisfactory percentage of the debt is repaid or a sufficient number of repayments has been made.

This paper introduces two modelling approaches to describe these patterns of repayment and non-repayment. The first is a payment sequence approach which looks at the movements at sequence level between a sequence of payments and a sequence of non-payments. The second is a survival analysis approach, which looks at whether there is a repayment or no repayment in each time period (usually a month). It models how many payments are made in a sequence until the debtor stops paying and how many missed periods occur before they start paying again. Using the average repayment rate per sequence for the first approach and the average repayment rate per period in the second approach, one can calculate the distribution of the overall repayment rate. The models are appropriate for portfolio level decisions and overall LGD rates. To estimate LGD for an individual, one needs to extend the models so the parameters are functions of the individual debtor's characteristics.

These approaches allow one to calculate the repayment rate under different write-off strategies as well as the average duration a debtor is in the collection process. This would allow the lender to decide on a suitable trade-off between the future recovery rate and the amount of future effort expended to reach that rate under the different write-off strategies. The results are relevant at the portfolio level since they involve the average recovery rate and the average extra effort involved. The models are not intended to identify the optimal write-off strategy but can be seen as a progress to optimising such decisions.

The next section gives an informal description of the data from the case study on which the models will be built. This is the type of data that collectors are now recording on a regular basis. Section 3 discusses the literature on collections processes as well as the use of Markov chain models in consumer lending. Section 4 describes the sequence based Markov chain model where the debtor moves between payment and non-payment sequences. Section 5 applies this model to the case study data to estimate recovery rate, and hence LGD, under simple write-off strategies. Section 6 describes a hazard rate model of the collections process. This involves more estimation than the sequence based model but allows much more complex write-off strategies. In both cases, a full model is outlined together with simplifications of the model which require fewer parameters as they assume that after an initial period the parameters of subsequent payment (and non-payment sequences) are the same. The models in Section 6 are applied to the case study data in Section 7. Finally conclusions are drawn from the models and their results.

2. Description of the Collections Data Set

The data we use in this case study describes the repayment history of 10,000 defaulted personal loans from a UK bank's loan book. These are loans that defaulted between 1988 and 1999 where default was defined as 90 days in arrears. The performance of the loans in the collection process was recorded from the start of 1988 until the end of 2003. The collections policy of the lender was to agree where possible with the debtor an amount that should be repaid each month until the debt was fully paid off. The data recorded whether there was a payment from the debtor in a given month and how much it was for. From that it is possible to see the history of sequences of payment and non-payment as shown in Figure 1.

[Figure 1 about here]

Figure 1 shows some examples of the actual payment patterns that occur in the data set we use later in the paper. The white bars are when the debtor is not repaying and the black bars are when the debtor is repaying. As can be seen from this graph, the debtors can go for long periods without paying and then start up again. In all payment patterns, the initial sequence must be of non-repayment since otherwise the debtor would not have been deemed to have defaulted. $NoPay_j$ is the j^{th} non-payment sequence and Pay_j is the j^{th} payment sequence. Some of the debtors never pay back anything after the default as for example debtor number 8. Others pay back part of their debt but are written off when they stop repaying, see for

ACCEPTED MANUSCRIPT

example debtor number 2. A third group pays back all of their debt – debtor number 4 for example, while others are still paying back at the end of the observation period. There are some debtors who are still repaying more than 120 months (10 years) after the default.

Recall that all debtors begin in $NoPay_1$ (their first non-payment sequence) since all of the debtors in the data set have defaulted. Note that during the time of this sample the definition of default started to be tightened to 90 days overdue. There are only two ways to leave $NoPay_1$. The debtor either has to start paying (Pay_1) or gets written off (W). Once the debtor starts paying there are only two ways to leave Pay_1. The debtor can either stop paying, in which case they enter $NoPay_2$ or they pay off all of their debt and are "cured" (C). In order to calculate the probability of a debtor entering Pay_j given that they are in $NoPay_j$, we take the number of those who reached $NoPay_j$ and divide that into the numbers who then enter Pay_j. This gives the probability $P(Pay_j|NoPay_j)$. Similarly the ratio of the number who reach $NoPay_{j+1}$ divided by those who reached Pay_j, gives the conditional probability $P(NoPay_{j+1}|Pay_j)$. These values together with the number $N(NoPay_i)$ and $N(Pay_i)$, which give the number of debtors in each relevant sequence, are given in Table 1. Since these are estimates of the probability of a Bernoulli random variable, the standard deviations, $(N(.)P(.)(1-P(.)))^{0.5}$ are also reported in Table 1.

Probabilities like $P(Pay_{j+}|NoPay_{j+})$ or $P(NoPay_{(j+1)+}|Pay_{j+})$ correspond to the calculation where we have taken the weighted average (weighted by number of cases) of the probabilities of the relevant transition for sequence j and all higher sequences. This is equivalent to assuming that all sequences later than the jth one have the same parameters. While there are debtors in the data set that continue on this stop/start payment process up to Pay_{25}, the proportion reaching $NoPay_{11}$ is less than 9%. Also, from Pay_4 and $NoPay_4$ onwards, the transition probabilities are getting quite close. This can be seen in Table 1 where the upper rows of the last column show the Chi-square test results to check if the proportion going from $NoPay_j$ to Pay_j is the same as that going from $NoPay_{(j+1)+}$ to $Pay_{(j+1)+}$. The lower section of the last column shows the results on the same test comparing the proportion moving from Pay_j to $NoPay_{j+1}$ is the same as that going from $Pay_{(j+1)+}$ to $NoPay_{(j+2)+}$. These results show that the parameters for moving from non-payment to payment sequences after the third payment sequence are similar. The probabilities of moving from payment to non-payment sequence are also converging if more slowly. To be conservative, in the full model we

ACCEPTED MANUSCRIPT

assume only that all payment and non-payment sequences after the tenth will have similar parameters to those of Pay_{10} and $NoPay_{10}$.

[Table 1 about here]

Table 2 describes the statistics of the total recovery rate distribution for the whole portfolio of defaulted loans under the lender's collection policy over the sample period.

[Table 2 about here]

The average recovery rate is 31.6 % corresponding to a LGD of 68.4% and the standard deviation was 29.2%. The form of the RR distribution is given by Figure 2 which has the U-shaped distribution common to almost all RR (and LGD) distributions. During the period of this sample, the collector had no fixed policy on writing off debts. Write off was done subjectively when it was felt the collections department was under pressure. It is clear from the examples in Figure 1 that such an approach allowed repayments to stop and start many times before the debts were written off.

[Figure 2 about here]

The collections data includes loans that defaulted in the period 1988-1994 and one that defaulted in 1995-1998. The first of these was an economic downturn and the second a period of recovery. Table 2 also looks at the differences in the recovery rate statistics for loans defaulting in these two periods. It appears that it is the economic situation in the collections process more than the economic situation at default which affects the collection results. Loans which default in the economic recession pay back a little more in the subsequent recovery than those which default during good economic times. This is seen in Table 2. It may be the case that those who default during a recession are more willing to try and repay when their economic situation improves than those who default in good times.

3. Literature Review

Consumer debt is a major factor in the current economic situation. As of the end of 2012, US consumers owed \$11.83 trillion with credit card debt being \$700 billion, student loan debt \$1.16 trillion and delinquency rates of 4.3% (Federal Reserve Bank of New York, 2015). In the UK, consumer debt stood at £1.445 trillion with £160.7 billion of this being unsecured credit (Bank of England, 2014). UK banks wrote off an average of £11.38 million of the defaulted portion of this debt each day in 2012 (Credit Action, 2013). Thus it is not surprising

ACCEPTED MANUSCRIPT

that there is an established literature on how consumers repay their loans (Kahlberg and Saunders, 1983). Perhaps what is surprising is how little attention has been paid in the literature on how consumers repay after they have defaulted.

One of the first Markov chain models of consumer credit card behaviour before default was suggested by Cyert, Davidson and Thompson (1962) and subsequent variations are reviewed in Thomas (2009). In a few instances, the payment pattern approach has also been used to rank borrowers in terms of their likelihood to default. It was used for instalment loans by Schwarz (2011) as a way of introducing new variables, namely the ratio of actual instalments payments made to those required. Stone (1976) used payment patterns to forecast when accounts receivable would be paid to a retail organisation. However, in that paper the whole cost must be paid off in one repayment. Stanford (1995) built an analytic solution to the accounts receivable forecast problem based on the Cyert-Davidson-Thompson model.

All these approaches have modelled the performance of borrowers before they have defaulted. Modelling payment patterns after default is different. The state space is not how overdue is the borrower's payment because all have already defaulted. Instead, it is whether the defaulter is currently in a repayment or non-repayment sequence. These payment patterns are modelled using a Markov chain, where the state space is whether the defaulter is currently paying (*Pay*), not paying (*NoPay*), Cured (*C*) where the whole defaulted amount has been repaid, or the loan has been written-off (*W*). Zhou (2011) is one of the few to consider the sequences of payment and non-payment in the collections process as a Markov chain. That thesis concentrated on only two aspects of the process. The first is the duration of the first non-payment sequence, i.e. the time until there is a first payment after default. The second is how likely the next repayment is likely to be severely late, i.e. that the non-payment sequences are above a certain duration. Our models concentrate on the whole collections process and so are able to estimate the average expected total recovery rate. Moreover, they allow one to look at the results under different write-off policies. The write-off decision does affect how much can be recovered in the collections process but has been ignored in most LGD models. Curnow et al. (1997) discuss the collection procedure at AT&T and Anderson (2007) has a more general discussion of the collections process but neither discusses the write-off decisions. An approach which looks at whether one should write off a loan is the dynamic programming model of the collections process found in de Almeida Filho et al. (2010). They concentrate on the optimal duration and sequence of the different collection actions that can be taken in the collections process. One of these is to cease collecting and so

ACCEPTED MANUSCRIPT

write off the loan. The state space in this model is the amount recovered so far, the current collection action and the duration so far of that action. It does not involve whether there was or was not a payment in the previous period which is the state space used in the models in this paper.

The models in this paper allow one to estimate the recovery rate (RR) and hence LGD. Basel II (BCBS, 2004) and Basel III (BCBS, 2011) banking regulations require that LGD be estimated for each segment of a bank's loan book. For corporate loans, there is a well-established literature on modelling LGD see the survey by Peter (2005) of the practical issues in LGD modelling. Most of the corporate LGD models are variants of a regression approach; see for example the papers in the book edited by Altman et al. (2005). More recent non-parametric variants, such as neural nets and regression trees, are compared in Loterman et al. (2011). Dermine and Neto de Carvalho (2006) investigated LGD for bank loans rather than for corporate bonds and showed how a log-log transformation led to a better regression fit of the data. Recently, Han and Jang (2013) have investigated how the lenders' actions can affect LGD for corporate credit. However, the literature on LGD for unsecured consumer loans is much more limited.

There are two methods of calculating LGD for the retail loans: workout LGD and implied historical LGD. Lucas (2006) suggested using the collection process to model LGD for mortgages. The collection process was split into whether the property was repossessed and the loss if there was repossession. A scorecard was built to estimate the probability of repossession and then a model used to estimate the "haircut" which is the percentage of the estimated sale value of the house that is actually realised at sale time.

Matuszyk et al. (2010) introduced a decision tree model for unsecured consumer loans to model the strategy of the collection process. This helps lenders decide whether to collect in-house, use an agent or sell off part of the defaulted loans. Our paper looks more at the operations one undertakes if one is collecting either in-house or as an agent. Bellotti and Crook (2012) added economic variables to the regression models estimating the LGD and found that their inclusion in the model was important. Zhang and Thomas (2012) examined whether it is better to estimate Recovery Rate (RR) or Recovery Amount. They used linear regression and survival analysis models to model Recovery Rate and Recovery Amount, so as to predict Loss Given Default (LGD) for unsecured personal loans. They found estimating Recovery Rate directly by using linear regression gave the best results.

In all the above quoted papers the results in terms of R-square values were poor - between 0.05 and 0.22. One reason for this is the lack of economic variables in the models. This could be addressed by using the dual time approach of Breeden (2007) which looks at vintage and maturity of the debt as well as economic conditions or by directly including economic variables into the regression (Bellotti and Crook, 2012) . Another reason is that the LGD distribution is far from normal and so regression approaches do not work without major modifications. A third reason why LGD is hard to predict is its dependence on the write-off policy the collector uses. The two models proposed hereafter give an alternative approach to modelling the recovery rate using payment patterns. These models have the advantage of including the write-off policy in the calculation and do not require the LGD distribution to have a specific form. The models could also be extended to include economic conditions.

The second model proposed here is a discrete time survival model. In other papers which use survival analysis in LGD modelling (Witzany et al, 2012; Bonini and Caivano, 2013), the time measured is directly the time until write-off. In this model the times measured are the lengths of the payment and non-payment sequences, which are then incorporated in the recovery rate estimate.

4. Modelling Repayment Patterns Using the Payment Sequence model

This model assumes the probability structure of payment and non-payment sequences is given by a Markov chain. Each sequence consists of one or more consecutive months of payment or non-payment. The recovery process always begins with a non-payment sequence, $NoPay_1$, since a borrower will only trigger the default by missing a payment. This is succeeded either by a payment sequence Pay_1 or a write-off, *W*. The payment sequence Pay_j either leads to complete repayment of all debt (*C*) or a further sequence of missed payments ($NoPay_{j+1}$) The process continues until either the loan is completely recovered (*C*) or written-off (W). See Figure 3. The Markov assumption means the number of payments or non-payments in a sequence does not affect the transition probabilities.

[Figure 3 about here]

The probabilities of the transitions are given by:

ACCEPTED MANUSCRIPT

$$P(Pay_j \mid NoPay_j) \text{ and } P(NoPay_{j+1} \mid Pay_j),\ j=1,2\ldots$$

Note that:

$$P(W \mid NoPay_j) = 1 - P(Pay_j \mid NoPay_j) \text{ and } P(C \mid Pay_j) = 1 - P(NoPay_{j+1} \mid Pay_j) \tag{1}$$

From this we are able to calculate the chance of being written off *by*:

$$P(W) = \sum_{i=1}^{\infty} P(W \mid NoPay_i) \prod_{1 \le j < i-1} P(NoPay_{j+1} \mid Pay_j) P(Pay_j \mid NoPay_j) \tag{2}$$

If we allowed the process of recovery to continue indefinitely, the chance of paying off all the debt must be:

$$P(C) = 1 - P(W) = \sum_{i=1}^{\infty} P(C \mid Pay_i) \prod_{1 \le j \le i} P(Pay_j \mid NoPay_j) \prod_{2 \le j \le i} P(NoPay_j \mid Pay_{j-1}) \tag{3}$$

It is unrealistic that the number of payments sequences be unlimited. The write off policy *WO(N)* writes off the debt at the start of the *(N+1)*$^{\text{th}}$ non-payment sequence. That would be the N^{th} time the debtor has stopped paying. In that case, the probability of full repayment is *P(C|N)* where:

$$P(C \mid N) = \sum_{i=1}^{N} P(C \mid Pay_i) \prod_{1 \le j \le i} P(Pay_j \mid NoPay_j) \prod_{2 \le j \le i} P(NoPay_j \mid Pay_{j-1}) \tag{4}$$

The probability of a write-off is then:

$$\begin{aligned} P(W \mid N) &= 1 - P(C \mid N) \\ &= \sum_{i=1}^{N} P(W \mid NoPay_i) \prod_{1 \le j < i} P(NoPay_{j+1} \mid Pay_j) P(Pay_j \mid NoPay_j) + \prod_{j=1}^{j=N} P(NoPay_{j+1} \mid Pay_j) P(Pay_j \mid NoPay_j) \end{aligned} \tag{5}$$

Zhang and Thomas (2012) showed that estimating recovery rates leads to more accurate models than estimating recovery amounts. So let *RR(i)* be the average recovery rate of the i^{th} payment sequence (the amount recovered in it as a fraction of the original defaulted amount). Note this is the average over those who have an i^{th} payment sequence but do not pay off all the loan in that sequence. For those who do pay off completely and so go $NP_i \rightarrow P_i \rightarrow C$, it is clear their recovery rate must be 1 when they reach *C*. So we add a recovery rate of $1 - \sum_{k=1}^{i} RR(k)$ when they reach C. With these estimates we can calculate the overall recovery rate (RR) if the lender does not write off any debt. Its expectation is *E(RR)* while if the recovery process is stopped at the *(N+1)*$^{\text{th}}$ non-payment sequence, the expected recovery rate is defined as *E(RR|N)*. These satisfy:

ACCEPTED MANUSCRIPT

$$\begin{aligned}E(RR) &= \sum_{i=1}^{\infty} RR(i)P(Pay_i \mid NoPay_i) \prod_{\min\{1,i-1\}\le j\le i-1} P(NoPay_{j+1} \mid Pay_j)P(Pay_j \mid NoPay_j) + \\ &\sum_{i=1}^{\infty} (\max(0,1-\sum_{k=1}^{k=i} RR(k)))P(C|Pay_i)P(Pay_i \mid NoPay_i) \prod_{\min\{1,i-1\}\le j\le i-1} P(NoPay_{j+1} \mid Pay_j)P(Pay_j \mid NoPay_j) \\ &= \sum_{i=1}^{\infty} \prod_{\min\{1,i-1\}\le j\le i-1} P(NoPay_{j+1} \mid Pay_j)P(Pay_j \mid NoPay_j)\left[P(Pay_i \mid NoPay_i)(RR(i)+(\max(0,1-\sum_{k=1}^{k=i} RR(k)))(1-P(NoPay_{i+1}|Pay_i)))\right]\end{aligned} \tag{6}$$

Similarly,

$$E(RR|\,N) = \sum_{i=1}^{N} \prod_{\min\{1,i-1\}\le j\le i-1} P(NoPay_{j+1} \mid Pay_j)P(Pay_j \mid NoPay_j) \left[P(Pay_i \mid NoPay_i)(RR(i)+(\max(0,1-\sum_{k=1}^{k=i} RR(k)))(1-P(NoPay_{i+1}|Pay_i)))\right] \tag{7}$$

This formulation assumes no interest is being charged on the defaulted debt, no discounting of the repayments and no collections costs. These are assumptions approved by some but not all regulators. One can modify the equation to deal with the first two of these and the third is dealt with in this paper by looking at the collection effort. *E(T)*, the expected number of payment sequences, is a good indicator of the effort and hence the cost involved in the collection process. Similarly, let *E(T|N)* be the expected number of payment sequences under policy *WO(N)*. Then:

$$E(T) = P(Pay_1 \mid NoPay_1)\left(1+\sum_{i=2}^{\infty}\prod_{j=2}^{i} P(Pay_j \mid NoPay_j)P(NoPay_j \mid Pay_{j-1})\right) \tag{8}$$

and $$E(T \mid N) = P(Pay_1 \mid NoPay_1)\left(1+\sum_{i=2}^{i=N}\prod_{j=2}^{i} P(Pay_j \mid NoPay_j)P(NoPay_j \mid Pay_{j-1})\right) \tag{9}$$

The lender will be aided in deciding which write-off policy to choose by comparing *E(RR|N)* with *E(T|N)* for different *N*. Alternatively they may look at the marginal reward per extra effort by looking at (*E(RR|N+1)-E(RR|N)*) / (*E(T|N+1)-E(T|N)*).

An advantage of this approach to estimating RR and LGD is that one gets a distribution of the recovery rates as well as the mean value. One measure of risk used in finance is Value at Risk (VaR). Estimate the α-quantile of the Recovery Rate, RR_α, i.e. the recovery rate if there is only an α chance of getting a worse recovery rate. Since the worst recovery occurs if the debtor does not leave $NoPay_1$, the second worst if it does not leave $NoPay_2$, and so on, we can estimate this by first defining N_α as the maximum *N*, so that:

ACCEPTED MANUSCRIPT

$$N_{\alpha} = Max\{N \mid \sum_{i=1}^{N} P(W \mid NoPay_i) \prod_{j<i} P(NoPay_{j+1} \mid Pay_j) P(Pay_j \mid NoPay_j) < \alpha\} \qquad (10)$$

Then the α-quartile of the recovery rate RR_α will be:

$$RR_{\alpha} = \sum_{i=Max\{1, N_{\alpha}-1\}}^{N_{\alpha}-1} RR(i) \qquad (11)$$

It seems reasonable to suppose the chance of a debtor paying for the first time is different from the chance someone who has already paid something but stopped paying will start to repay again. However, it would seem reasonable that for someone in this latter position it would not matter too much after a time how many payment sequences have already occurred or what their condition was when they defaulted. Similarly, the chance that a defaulter who has started to pay for the first time, stops paying before paying off the whole amount is likely to be different to the chance that someone who has already paid something and then stopped, but is now paying again, stops again before paying off. Again for debtors in this latter position, after a time it will not matter too much how many times they have previously stopped paying. Similarly, the recovery rate in the first sequence might be different to that in the second sequence. However, one might expect the recovery rates in the fifth, sixth and higher sequences to be very similar. (This is for those who do not cure in that sequence).

Assume that the first K payment and non-payment sequences are different from one another but all subsequent sequences of payment have the same probabilities of stopping and the same recovery rate estimates. Similarly, assume all non-payment sequences from the K^{th} have the same probabilities of a subsequent payoff occurring. Then one only needs to estimate 2K probabilities and K recovery rates. This would lead to the parameters:

$$\begin{aligned} &P(Pay_i \mid NoPay_i) = p_i, 1 \le i \le K-1; P(Pay_i \mid NoPay_i) = p_K, \forall i \ge K; \\ &P(NoPay_{i+1} \mid Pay_i) = q_i, 1 \le i \le K-1; P(NoPay_{i+1} \mid Pay_i) = q_K, \forall i \ge K; \\ &RR(i) = r_i, 1 \le i \le K-1; RR(i) = r_K, \forall i \ge K; \end{aligned} \qquad (12)$$

In that case, the equations (2), (5), (6), (7), (8) and (9) reduce to the following:

$$\mathrm{P(W)=(1\text{-}p_1)} + \sum_{i=2}^{i=K-1} (1-p_i) \prod_{j=1}^{j=i-1} p_j q_j + \frac{1-p_K}{1-p_K q_K} \prod_{j=1}^{J=K-1} p_j q_j \qquad (13)$$

$$\mathrm{P(W|N)=(1\text{-}p_1)} + \sum_{i=2}^{i=\min\{K,N\}} (1-p_i)\prod_{j=1}^{j=i-1} p_j q_j + \delta(K-N+1)\prod_{j=1}^{j=N} p_j q_j$$

$$+\delta(N-K)\prod_{j=1}^{K} p_j q_j \left(\frac{1-p_K}{1-p_K q_K}\left(1-(p_K q_K)^{N-K}\right) + (p_K q_K)^{N+1-K} \right) \qquad (14)$$

where δ(Y)=1 if Y>0; δ(Y)=0 if Y $\le$ 0

$$E(RR) = r_1 p_1 + \sum_{i=2}^{K-1} r_i p_i \prod_{j=1}^{i-1} p_j q_j + \prod_{j=1}^{j=K-1} p_j q_j \left(r_K p_K \frac{p_K q_K}{1-p_K q_K} \right) +$$

$$\sum_{i=1}^{K-1} p_i (1-q_i)(\max\{0, 1-\sum_{s=1}^{s=i} r_s\}) \prod_{j=1}^{j=i-1} p_j q_j + (\prod_{j=1}^{j=K-1} p_j q_j) p_K (1-q_K) \sum_{i=1}^{\infty} \left(p_K q_K \right)^{i-1} \left(\max\{0, 1-\sum_{s=1}^{K-1} r_s - i r_K\} \right) \qquad (15)$$

$$E(RR \mid N) = r_1 p_1 + \sum_{i=2}^{\min\{K,N\}} r_i p_i \prod_{j=1}^{i-1} p_j q_j + (1-r_1) p_1 (1-q_1) + \sum_{i=2}^{\min\{K,N\}} (\max\{0, 1-\sum_{s=1}^{s=i} r_s\})(1-q_i) p_i \prod_{j=1}^{i-1} p_j q_j$$

$$+\delta(N-K)\left(\prod_{j=1}^{j=K} p_j q_j \left(r_K p_K \frac{1-(p_K q_K)^{N-K}}{1-p_K q_K} \right) + \sum_{i=1}^{K-N} (\max\{0, 1-\sum_{s=1}^{s=K} r_s - i r_K\})(p_K q_K)^i (1-q_K) p_K \right) \qquad (16)$$

$$E(T) = p_1 + \sum_{i=2}^{K-1} p_i \prod_{j=1}^{i-1} p_j q_j + \prod_{j=1}^{j=K} p_j q_j \left(\frac{p_K}{1-p_K q_K} \right) \qquad (17)$$

$$E(T \mid N) = p_1 + \sum_{i=2}^{\min\{K,N\}} p_i \prod_{j=1}^{i-1} p_j q_j + \partial(N-K) \prod_{j=1}^{j=K} p_j q_j \left(p_K \frac{1-(p_K q_K)^{N-K}}{1-p_K q_K} \right) \qquad (18)$$

These are the formulae which we will use in the case study calculations in section 5. One might think that if $N>K$ one would want to make the same decision of whether to carry on or write off no matter what the value of N. However, this is not the case because the recovery rate of the final payment when a debtor cures lessens as the number of previous payment sequences increases. Thus there comes a time when it is worth writing the debt off even though that might not have been the case when $N=K$.

5. Case Study using the payment Sequence Model on the collections data

ACCEPTED MANUSCRIPT

The case study uses the data set described in section 2. Table 1 of that section gives the transition probabilities between payment and non-payment sequences needed for the modelling. Table 3 describes the repayment rates per sequence *RR(i)* from the data.

[Table 3 about here]

RR(i) is calculated by taking the average of the repayment rate (the amount of repayment in the sequence as a ratio of the original debt) for the i^{th} payment sequences in which the borrower stops paying before having completely paid off the debt. These repayment rates start with a rate of 13.15% in the first payment sequence and drop monotonically until the value is 5.91% in the tenth sequence. Although this repayment rate is always dropping, the values are slowly converging and so assuming a constant *RR(i)* after the tenth payment sequence is a reasonable assumption. The last five entries in Table 1 are the average repayment rates if we combine all the repayment sequences from the fifth, fourth, third, second and first onwards. Recall the last column of Table 1 looks at the chi-square tests results for the hypotheses that $P(Pay_i|NoPay_i)$ and $P(Pay_{i(i+1)+}|NoPay_{(i+1)+})$ take the same value and also the same thing for the hypothesis $P(NoPay_{i+1)}|Pay_i)$ equals $P(NoPay_{i(i+2)+}|Pay_{(i+1)+})$. Table 3 reports the results for *RR(i)* $i=1,\ldots 10$ and *RR(i+)*, $i=1,..,5$. It suggests one can use the same parameters for all sequences from the third and probably the second onwards. The difference between *RR(i)* and *RR(i+1)* for such sequences is less than 0.01 and getting smaller as *RR(i)* converges to 0.59.

Substituting the values in Tables 1 and 3 into equations (14) to (17) gives the expected average total recovery rate and the average expected number of payment periods involved under a number of write-off policies.

[Table 4 about here]

Table 4 gives the results for $E(RR|N)$ and $E(T|N)$ for different write-off policies, *WO(N)*, where one is writing off the debt when the borrower stops paying for the N^{th} time.. The final row corresponds to the values if the collector never wrote off any debtor. In that case, the expected total recovery rate is 38.2%, corresponding to a LGD of 61.8%. The first ten rows show the results under the write-off policies of writing off any borrower when they reach the $N+1^{th}$ non-payment sequence (i.e. the borrower has stopped paying *N* times). Note if one stops the process after the tenth time a borrower stops paying, the lender will expect to recover 37.1% of the debt. If one wants to recover at least 30% of the debt on average, one should write off the debt on the fifth time the borrower stops a payment sequence. A harsher policy of two failures and the borrower is written off leads to an expected recovery rate of 18.0%. Comparing tables 2 and 4 shows that the collector's current recovery rate of 31.6% could be obtained by writing off debts the 6^{th} time a debtor stops paying.

[Figure 4 about here]

Figure 4 gives a graphical representation of the trade-off between the total expected recovery rate and the total number of expected payment periods under these different write-off policies. The graph is concave which means the increase in recovery rate per extra payment period decreases as the write-off policy increases in *N*. To find the recovery rate per payment sequence, calculate the slope from the origin to the point corresponding to that policy on the curve. The best result in terms of recovery rate per payment sequence is to write off the debt after the first non-payment but this leads to a low recovery rate of 10.7%. At the other end of the graph, the difference between the policy of no active write-offs and writing off at the tenth non-payment is an increase in expected recovery rate of 1.1% but only an expected increase of 0.266 in the number of payment sequences. So although one gets very little recovery deep into the collections process, it also does not involve much more effort because few debtors have had such a large number of payment sequences.

If one assumes the ratio of the average defaulted amount to the cost of keeping a debtor in the collections process for one more non-payment-payment cycle is 10, then from Figure 4 we can see the optimal write-off policy is when the tangent to the curve has a slope of 10, i.e. at (2.639, 0.331). The same result can be obtained from Table 4 by maximizing $10E(RR|N)-E(T|N)$ which happens at $N=6$ with $10(0.331)-2.639=0.671$.

To see the effect of taking the simpler models where one assumes that all payment and non-payment sequences after the K^{th} are given the same parameters, we undertake the calculation for the expected recovery rate with this assumption holding for $K=5,4,3,2$ and 1. Recall that the chi-square tests suggest $K=2$ or 3 are sensible choices. The previous calculation assumed that after the 10^{th} sequence all the remaining sequences have the same parameter. The difference between the recovery rate using this assumption and the simplified assumptions that assume similarity after the first, second, third, fourth and fifth sequence for all future sequences is shown in Table 5.

ACCEPTED MANUSCRIPT

[Table 5 about here]

In that table, a positive value says the simplifying assumption has come up with a lower recovery rate, while a negative value means it has resulted in a higher recovery rate. The maximum error using the 1+ simplification, which means all the sequences have the same parameters, is 2.36%. For 2+, where the first sequence is assumed different from all the rest, it is 1.51%. For 3+, 4+ and 5+, the maximum errors are just below or above 1%. This suggests it is enough to make *K=2* or *K=3* (which involves estimating 6 and 9 parameters respectively) to get an accurate model. This table is useful in showing the differences between these simplified models with 6 or 9 parameters and a larger one, which in this *K=10* case has 30 parameters. It is fairly obvious though from the results in the last column of Table 5 that one should at least use a model which differentiates between the first payment and non-payment sequence and the rest.

6. Modelling the Recovery Process Using a Hazard Rate Model

In this section, we develop a hazard rate model which requires more parameters but can evaluate the impact of more sophisticated write-off policies than the payment sequence model. The model estimates the likelihood of transition from payment to non-payment (or vice versa) each month. This allows the duration of each payment sequence to be modelled. By adding data about the repayment rate in each month, estimates of the total repayment rate can be made.

We extend the notation introduced earlier by defining $NoPay_i^j$ to be the j^{th} period of non-payment in a non-payment sequence i and let Pay_i^j be the j^{th} period of payment in payment sequence i. The state space of the system now extends to that shown in Figure 5.

[Figure 5 about here]

All defaulters start with a non-payment month which is labelled $NoPay_1^1$. The process can then move to one of three states:

- $NoPay_1^2$ when there is no payment and so the non-payment sequence continues
- Pay_1^1 if there was a payment and the first payment sequence starts
- *W* where the debt is written off and the recovery action ceases.

For a month where there is a payment, say Pay_i^j, the process can again move to three different states, namely:

ACCEPTED MANUSCRIPT

- Pay_i^{j+1} if there is a payment in the next month so the payment sequence continues
- $NoPay_{i+1}^1$ if there is no payment and so a new non-payment sequence begins
- C where the payment is enough to pay off all the defaulted amount and so the loan is "cured".

The conditional probabilities $P(Pay_i^1 | NoPay_i^j)$ and $P(NoPay^1_{i+1} | Pay_i^j)$ can be thought of as discrete time hazard functions which determine respectively how long a non-payment and a payment sequence lasts.

Let RT be the expected total recovery rate to date in the process and define $RRM(i)$ to be the average recovery rate paid per month in payment sequence i. Whenever the system moves to Pay_i^j for any j, $RRM(i)$ is added to RT. If this means that RT becomes greater or equal to 1, then the process moves to state C. So it is the value of the variable RT that determines when the process enters the cure state. The model could be extended by making the average monthly repayment $RRM(i,j)$ to be a function of how long the repayment sequence has been as well as the number of previous sequences. This would allow the situation where there is a large payment made at the start of each repayment sequence.

This model can deal with more write-off policies. Define $WO(N,M)$ to be the policy that writes off either after the N^{th} time the debtor stops a payment sequence, or when it is M or more periods since the collection process started and there is a non-payment this period. The first condition occurs when the process reaches the state $NoPay_{N+1}^1$, while the second condition requires the state of the process to include the number of periods since the start of the collection process. Thus the states of the system are (Pay_i^j, RT, m) or $(NoPay_i^j, RT, m)$, where Pay_i^j ($NoPay_i^j$) denote the collection process is in the jth period of the ith payment (non-payment) sequence, respectively. RT is the recovery rate so far and m is the number of periods since the start of the collection process.

The transition between states is given as follows:

- $(Pay_i^j, RT, m) \to (Pay_i^{j+1}, RT + RRM(i), m+1)$ with transition probability $P(Pay_i^{j+1} | Pay_i^j)$ provided $RT + RRM(i) < 1$
- $(Pay_i^j, RT, m) \to (C, 1, m+1)$ with transition probability $P(Pay_i^{j+1} | Pay_i^j)$ provided $RT + RRM(i) \geq 1$

- $(Pay_i^j, RT, m) \rightarrow (NoPay_{i+1}^1, RT, m+1)$ with transition probability $P(NoPay_{i+1}^1 \mid Pay_i^j)$ provided $m+1 < M$ and $i < N$
- $(Pay_i^j, RT, m) \rightarrow (W, RT, m+1)$ with transition probability $P(NoPay_{i+1}^1 \mid Pay_i^j)$ provided $m+1 = M$ or $i = N$
- $(NoPay_i^j, RT, m) \rightarrow (Pay_i^1, RT + RRM(i), m+1)$ with transition probability $P(Pay_i^1 \mid NoPay_i^j)$ provided $RT + RRP(i) < 1$
- $(NoPay_i^j, RT, m) \rightarrow (C, 1, m+1)$ with transition probability $P(Pay_i^1 \mid NPay_i^j)$ provided $RT + RRM(i) \geq 1$
- $(NoPay_i^j, RT, m) \rightarrow (NoPay_i^{j+1}, RT, m+1)$ with transition probability $P(NoPay_i^{j+1} \mid NoPay_i^j)$ provided $m+1 < M$
- $(NoPay_i^j, RT, m) \rightarrow (W, RT, m+1)$ with transition probability $P(NoPay_i^{j+1} \mid NoPay_i^j)$ provided $m+1 \geq M$

The expected total recovery rate under such write-off policies is calculated by an iterative scheme beginning with $m=M$ and then working back through the states in decreasing order of m value. Eventually one can calculate the overall total expected recovery rate, $E(RR)$, which is that in the first state of the sequence, namely $E_{RR}(NoPay^1{}_1, 0, 0)$. For the write-off policy $WO(N,M)$, this means solving the following set of equations (20) with the boundary conditions as in (19):

$$
\begin{aligned}
&E_{RR}(Pay_i^j, RT, m) = 1 \quad \text{if } RT \geq 1;\\
&E_{RR}(NoPay_{N+1}^1, RT, m) = RT;\\
&E_{RR}(NoPay_i^j, RT, M) = RT
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
&E_{RR}(Pay_i^j, RT, m)\\
&\quad = P(Pay_i^{j+1} | Pay_i^j) E_{RR}(Pay_i^{j+1}, RT + RRM(i), m+1) +\\
&\quad P(\mathrm{No}Pay_{i+1}^1 | P_i^j) E_{RR}(Pay_{i+1}^1, RT, m+1)\\
&E_{RR}(NoPay_i^j, RT, m) =\\
&\quad P(Pay_i^1 | NoPay_i^j) E_{RR}(Pay_i^1, RT + RRM(i), m+1) +\\
&\quad P(\mathrm{No}Pay_i^{j+1} | NoPay_i^j) E_{RR}(NoPay_i^{j+1}, RT, m+1)
\end{aligned} \tag{20}
$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Define $T(Pay_i^j, RT, m)$ (or $T(NoPay_i^j, RT, m)$) to be the time in the collection process given one is in state (Pay_i^j, RT, m) (or $(NoPay_i^j, RT, m)$) and $E(T)(Pay_i^j, RT, m)$ and $E(T)(NoPay_i^j, RT, m)$ being their expected values. These are calculated from an identical set of equations to (19) and (20) but with slightly different boundary conditions namely:

$$E(T)\left(Pay_i^j, RT, m\right) = m \text{ if } RT \geq 1;$$
$$E(T)(NoPay_{N+1}^1, RT, m) = m; \quad E(T)\left(NoPay_i^j, RT, M\right) = M \qquad (21)$$

$$\begin{aligned} E(T)(Pay_i^j, RT, m) = {} & P(Pay_i^{j+1}|Pay_i^j)E(T)\left(Pay_i^{j+1}, RT + RRM(i), m + 1\right) + \\ & P(NoPay_{i+1}^1|Pay_i^j)E(T)(Pay_{i+1}^1, RT, m + 1) \\ E(T)(NoPay_i^j, RT, m) = {} & P(Pay_i^1|NoPay_i^j)E(T)(Pay_i^1, RT + RRM(i), m + 1) + \\ & P(NoPay_i^{j+1}|NoPay_i^j)E(T)\left(NoPay_i^{j+1}, RT, m + 1\right) \qquad (22) \end{aligned}$$

These equations calculate the expected average recovery rate and the average number of periods that the collection process takes under different write-off policies. This allows management to decide on what is the appropriate policy for them. There is a trade-off between increasing the recovery rate and increasing the time and hence the effort and cost of the collection process. As before, if the average default amount is *b* and the cost each period a debtor is in the collection process is *c*, then the collector can find the expected profitability of each strategy by calculating $bE(RR) - cE(T)$. In this way, one can calculate the most profitable write-off policy.

7. Case Study: Applying Hazard Rate Model to In-house Collections Data

The hazard rate model of the previous section is now applied using the collections data described in Section 2. This model involves estimating the probabilities, $P(Pay_i^{j+1}|Pay_i^j)$ and $P(NoPay_i^{j+1}|NoPay_i^j)$. The number of probabilities to be estimated can be limited by assuming all payment and non-payment sequences after the K^{th} ones have the same probability parameters as the K^{th} one. Even then, there are theoretically an infinite number of probabilities as *j* can take a countable number of values. So we add the assumption that in every sequence the hazard rates $P(Pay_i^{j+1}|Pay_i^j)$ and $P(NoPay_i^{j+1}|NoPay_i^j)$ are constant once $j \geq L$. This seems a reasonable assumption and is in most cases supported by the chi-square test results in Table 6. Also, once a debtor has settled into a payment or non-payment sequence, it appears not to matter how long the sequence has already lasted. With these assumptions, one is left with *2LK* different possible states, Pay_i^j and $NoPay_i^j$. In our case, we take *K=L=3*

[Table 6 about here]

An alternative approach to estimating $P(Pay_i^{j+1} | Pay_i^j)$ for all *j* would be via a parametric hazard rate functions $h_{P_i}(.)$ where $h_{P_i}(j) = P(Pay_i^{j+1} | Pay_i^j)$ and $h_{NP_i}(j) = P(NoPay_i^{j+1} | NoPay_i^j)$. This would avoid putting a limit L on the non-constant part of each hazard rate function but computational requirements would still require that the conditional probabilities be constant for *j* where $j \geq L$. In this paper we will use the actual conditional probabilities rather than the hazard rate function but with *K=L=3*. The resultant 18 probabilities are given in Table 6.

Initially RT looks continuous as it could take any value in [0,1]. However if all sequences after K^{th} repayment sequence have the same repayment rates per period, RT will only take a discrete number of values. In this case, the only repayment rates per period that need be considered are *RRM(1)*, *RRM(2)*, …, *RRM(K)*. Define:

$$r=gcd\{RRM(1), RRM(2), \ldots, RRM(K)\} \text{ and } RRM(i)= rd(i)$$

We can then replace *RT* in any state by integers *d=0,1,...,D*, where $RT = rd$ and $D = [1/r]+1$. In this hazard model, there is no extra term to deal with the cases when a repayment means the whole amount is repaid and the collection is complete. Thus, unlike the first model, we include all the cases which involve an *i*th repayment sequence when calculating the *RRM(i)* values in Table 8.

[Table 7 about here]

There are *2LKD* possible states, which in this case with *K=L=3*, *D=541* leads to 9738 states. If we denote the states (Pay_i^j, rd, m) as $(1, i, j, d, m)$ and $(NoPay_i^j, rd, m)$ as $(0, i, j, d, m)$, we can calculate the expected recovery rates under a $WO(N, M)$ write-off policy by solving (19) and (20) in the equivalent form.

$$E_{RR}(1, i, j, D, m) = 1; E_{RR}(0,1, N+1, d, m) = rd; E_{RR}(0, I, j, d, M) = rd;$$
$$E_{RR}(1, i, j, d, m) = P(P_i^{j+1} | P_i^j) E_{RR}(1, i, \min\{j+1, L\}, d + d(i), m+1) +$$

$$P(NP_{i+1}^{1}|P_i^j)E_{RR}(0, i+1, 1, d, m+1);$$

$$E_{RR}(0, i, j, d, m) = P(NP_i^{j+1}|NP_i^j)E_{RR}(0, i, \min\{j+1, L\}, d, m+1) + P(P_i^1|NP_i^j)E_{RR}(1, i, 1, d+d(i), m+1)$$

[Table 8 about here]

Table 8 shows the results of these calculations. If one compares the results of the final column of Table 8, E(RR) with *M=48*, with the *E(TT|N)* column in the first three rows and the last row of Table 4, one finds that the two models come up with quite similar results. Even though the last column in Table 8 corresponds to the collection process stopping after four years while that of Table 4 has no limit on how long the process lasts, the recovery rates vary by at most 2%. Apart from the *N*=1 case, the Table 8 results with their fixed time limits are less than the Table 4 results with their collection process of unlimited duration. A comparison with the actual expected recovery rate of 31.6% would suggest a write off policy around *N=3*, *M=61* or one of *N=∞, M=43*.

[Figure 6 about here]

As Figure 6 shows, increasing *N*, the number of payment sequences before write-off, by 1 has less effect on the recovery rate than increasing *M*, the duration of the collections process by 12 (months). The graph for each *N* has an initial convex part, followed by an almost linear section and finally a concave section. In the case of *N*=∞, the curve is convex until *M=54* and concave thereafter. One can think of this as an initial learning process, where defaulters overcome the initial reluctance to repay; then a steady state; and finally an ageing process as the defaulters who are left are the ones who are least likely to repay. The change from convexity to concavity varies depending on *N*. For *N=1*, it occurs at *M=20*; for *N=2* at *M=28* and for *N=3* at *M=31*. The graph shows though what impact the write-off policies have in terms of recovery rates. Even comparing two realistic policies (*N*=3, *M*=34) and (*N*=∞, *M*=48), the recovery rate varies from 11% to almost 37%. Moreover, comparing the *N*=∞ case with the actual recovery rate in Table 2 shows that the current policy pursued has the same recovery rate as one where debtors are only written off after 48 month since the collection process. The *N*=∞ leads to high recovery rates after 8 years but this is such an unrealistic policy (no write-off for 8 years no matter what) that there is a lack of data available. One can find the most profitable write-off policy of this form by repeating what was done in section 4 and calculating $E(T|N,M)$ the expected time in the collections process under the $WO(N,M)$ write-off policy. This will be less than M because some of the borrowers will pay off before then or be written off under the N part of the rule. If the average defaulted amount is a and the cost of collections per period per debtor is c, then one should maximise $aE(RR|N,M)-cE(T|N,M)$ as was the case in section 5.

8. Conclusions

The paper discusses a way of modelling recovery rates (RR) and hence Loss Given Default (LGD) since $LGD=1-RR$ for unsecured consumer loans. It models the patterns of how debtors pay back their debt after default. These models not only predict loss given default but they highlight how LGD values depend on the write off policy of the collector. By allowing collectors to estimate both the extra proportion recovered and the extra effort involved if a write-off policy is relaxed, they indicate what write-off strategies are optimal.

There are two related models developed in this paper. The first uses the sequences of consecutive payments or consecutive non-payments as the basic units together with the average recovery rate in each payment sequence. Markov chain ideas lead to an overall recovery rate model. The second model uses a discrete hazard rate approach to estimate the chance a defaulter is paying or not paying in a given month. Such a model gives estimates of the duration of each payment or non-payment sequence. Estimating the monthly average recovery rate allows one to estimate the total recovery rate.

The first model requires less data to implement and leads to an analytic solution. However the write-off policies one can consider with it are somewhat limited. The second approach involves more parameters and has to be solved iteratively. However it can deal with more complex write-off policies. The write-off policies considered in this paper depend on the current duration of the collections process and the number of times a borrower fails to continue in a payment sequence. The model could also deal with write-off policies which depended on what percentage of the debt had been recovered, the time since the last payment, as well as combinations of these four elements.

These models are essentially work in progress and indicate what is possible with this repayment pattern approach. The models operate at the portfolio level, since they deal with average recovery rates and average time in the collections process. They are useful as

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

operations management tools for the collections process. They help the collector decide what the optimal trade-off is between recovery rate and time, effort and cost in the collections process. This leads to decisions on the staffing levels and skills needed for collections.

One could extend the models to work at the individual defaulter level. The parameters would then be functions of the debtor's characteristics and the prior performance in collections. This develops the idea of a collection score suggested in Anderson (2007). The models could easily be extended to allow for discounting of the later repayments, and introducing collection costs explicitly. The models in this paper calculate the expected recovery rate and the expected collections effort and allows the lender to determine the appropriate trade-off, rather than requiring explicit collection costs. One could also introduce economic conditions by making the recovery rates and the transition probabilities be functions of economic variables. This would improve the accuracy of the models but must wait until data on collections performance against economic conditions is regularly collected. Thus though these models are some of the first to take a repayment pattern approach to LGD and RR modelling, there are clear indications of how to develop this approach. It has the advantage that it does not depend on the form of the LGD distribution, is able to deal with collector's operating decisions, such as their write-off policy, and could include economic effects. These are three of the issues that cause difficulties in LGD modelling.

References

Altman, E., Resti, A. & Sironi, A. (2005). *Recovery Risk: The next Challenge in Consumer Risk Management*. Risk Books, London.

Anderson, R. (2007). *The Credit Scoring Toolkit Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.

Bank of England (2014), Statistical Release, Money and Credit, May 2014, London.

Basel Committee on Banking Supervision (BCBS) (2004). International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Basel Committee on Banking Supervision, Bank for International Settlements.

Basel Committee on Banking Supervision (BCBS) (2011). Basel III: a Global Regulatory Framework for More Resilient Banks and Banking Systems. Basel Committee on Banking Supervision, Bank for International Settlements.

ACCEPTED MANUSCRIPT

Bellotti, T. & Crook, J.N. (2012). Loss Given Default Models Incorporating Macroeconomic Variables for Credit Cards. *International Journal of Forecasting*, *28*, 171-182.

Bonini s. & Caivano G. (2013) The survival analysis approach in Basel II credit risk management: modelling danger rates in the loss given default parameter. *Journal of Credit Risk*, 9(1), 101-118.

Breeden, J.L., (2007) Modeling data with multiple time dimensions. *Computational Statistics& Data Analysis*, 51,4761 – 4785.

Credit Action (2013). UK Debt Statistics March 2013 http://www.creditaction.org.uk/helpful-resources/debt-statistics.html

Curnow, G., Kochman, G., Meester, S., Sarkar, D. & Wilton, K. (1997). Automating Credit and Collections Decisions at AT&T Capital Corporation. *Interfaces*, *27*, 29-52.

Cyert, R.M., Davidson, H.J. & Thompson, G.L. (1962). Estimation of Allowance for Doubtful Accounts by Markov Chain. *Management Science*, *8*, 287-303.

De Almeida Filho, A. T., C. Mues, Thomas L.C. (2010) Optimizing the Collections Process in Consumer Credit, *Production and Operations Management* **19**(6): 698-708.

Dermine, J. & Neto de Carvalho, C. (2006). Bank Loan Losses-Given-Default: a Case Study. *Journal of Banking & Finance*, *30*, 1219-1243.

Federal Reserve Bank of New York (2015). Household Debt and Credit Report. February 2015. http://www.newyorkfed.org/householdcredit/

Han, C. & Jang, Y. (2013). Effects of Debt Collection Practices on Loss Given Default. *Journal of Banking & Finance*, *37*, 21-31.

Kallberg, J.G. & Saunders, A. (1983). Markov Chain Approaches to the Analysis of Payment Behavior of Retail Credit Customers. *Financial Management*, *12*, 5-14.

Lucas, A. (2006). Basel II Problem Solving, QFRMC Workshop and Conference on Basel II & Credit Risk Modelling in Consumer Lending. University of Southampton, Southampton, U.K.

Loterman, G., Brown, I., Martens, D., Mues, C. & Baesens, B. (2011). Benchmarking Regression Algorithms for Loss Given Default Modelling. *International Journal of Forecasting*, *28*,161-170

ACCEPTED MANUSCRIPT

Matuszyk, A., Mues, C. & Thomas, L.C. (2010). Modelling LGD for Unsecured Personal Loans: Decision Tree Approach. *Journal of Operational Research Society, 61*, 393-398.

Peter, C. (2005), Estimating Loss Given Default – Experiences from Banking Practice, pp 143-175, in: the Basel II Risk Parameters, Ed Englemann B., Rauhmeier R., Springer, Berlin.

Schwarz, A. (2011). Measurement, Monitoring and Forecasting of Consumer Credit Default Risk An Indicator Approach Based on Individual Payment Histories, German Institute for International Educational Research Frankfurt am Main, Germany, Schumpeter Discussion Papers 2011-004.

Stanford, R.E. (1995). A Structured Sensitivity Analysis for a Markov Model of Accounts Receivable. *Journal of Accounting, Auditing & Finance*, *10*, 643-65.

Stone, B.K. (1976). The Payments-Pattern Approach to the Forecasting and Control of Accounts Receivable. *Financial Management*, *5*, 65-82.

Thomas, L.C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press, Oxford.

Witzany, J., Rychnovsky, M. & Charamza P., (2012), Survival analysis in LGD modeling. *European Financial Accounting Journal*, 2012(1), 7-27.

Zhang, J. & Thomas, L.C. (2012). Comparisons of Linear Regression and Survival Analysis Using Single and Mixture Distributions Approaches in Modelling LGD. *International Journal of Forecasting*, *28*, 204-215.

Zhou, F. (2011). Event History Analysis for Debt Collection Portfolio. Ph.D. Thesis, Imperial College London, U.K.

ACCEPTED MANUSCRIPT

i seq. no.	N(NoPay$_i$)	N(Pay$_i$)	P (Pay$_i$| NoPay$_i$)	S.D. P(Payi|No Payi)	P (W|NP$_i$)	P (NoPay(i +1)| Payi)	S.D. P(C|Pa yi)	P (C|Pay$_i$)	Segments compared	Chi-square value
1	9998	7180	0.718	0.004	0.282	0.980	0.002	0.020	**P(Pay|NoPay)**	
2	7036	5632	0.800	0.005	0.200	0.973	0.002	0.027	4v5+	5.528
3	5482	4524	0.825	0.005	0.175	0.967	0.003	0.033	3v4+	5.695
4	4374	3719	0.850	0.005	0.150	0.961	0.003	0.039	2v3+	46.74
5	3575	2960	0.828	0.006	0.172	0.955	0.004	0.045	1v2+	559.47
6	2826	2369	0.838	0.007	0.162	0.954	0.004	0.046		
7	2260	1917	0.848	0.008	0.152	0.957	0.005	0.043		
8	1834	1560	0.851	0.008	0.149	0.940	0.006	0.060	**P(NoPay|Pay)**	
9	1466	1214	0.828	0.010	0.172	0.921	0.008	0.079	4v5+	114.15
10	1118	903	0.808	0.012	0.192	0.924	0.009	0.076	3v4+	229.97
5+	13079	10923	0.835	0.003	0.165	0.946	0.002	0.054	2v3+	246.63
4+	17453	14642	0.839	0.003	0.161	0.950	0.002	0.050	1v2+	287.08
3+	22935	19166	0.836	0.002	0.164	0.954	0.002	0.046		
2+	29971	24798	0.827	0.002	0.173	0.959	0.001	0.041		
1+	39969	31978	0.800	0.002	0.200	0.963	0.001	0.037		

Table 1: Transition probabilities between payment sequences from case study data

Year	N	Mean	S.D.
All	9998	31.6%	29.2%
1988-1994	5645	32.3%	28.9%
1995-1999	4353	30.5%	29.5%

Table 2: Recovery rate statistics for full case study portfolio

i: sequence number	***N(Pay$_i$)***	**RR(i)**
1	7180	0.1315
2	5632	0.1095
3	4524	0.0971
4	3719	0.0908
5	2960	0.0846
6	2369	0.0793
7	1917	0.0738
8	1560	0.0687
9	1214	0.0638
10+	903	0.0591

i*+: sequences starting from *i	***N(Pay$_i$)***	**RR(i)**
1+	31978	0.0987
2+	24798	0.0892
3+	19166	0.0832
4+	14642	0.0789
5+	10923	0.0749

Table 3. Repayment rates per sequence from case study data

ACCEPTED MANUSCRIPT

N	E(RR|N)	E(T|N)
1	10.7%	0.718
2	18.0%	1.281
3	23.4%	1.734
4	27.6%	2.106
5	30.7%	2.402
6	33.1%	2.639
7	34.8%	2.831
8	35.8%	2.987
9	36.6%	3.108
10	37.1%	3.198
∞	38.2%	3.464

Table 4. Recovery rate and number of payment sequences under different write-off policies *WO(N)* where *N* is number of occasions borrower stopped repaying

WO(N)	(E(RR|N) with 5+) - (E(RR|N) with 10+)	(E(RR|N) with 4+) - (E(RR|N) with 10+)	(E(RR|N) with 3+) - (E(RR|N) with 10+)	(E(RR|N) with 2+) - (E(RR|N) with 10+)	(E(RR|N) with 1+) - (E(RR|N) with 10+)
1	0.0000	0.0000	0.0000	0.0000	-0.0236
2	0.0000	0.0000	0.0000	-0.0115	-0.0233
3	0.0000	0.0000	-0.0058	-0.0141	-0.0195
4	0.0000	-0.0044	-0.0092	-0.0151	-0.0168
5	-0.0027	-0.0065	-0.0102	-0.0143	-0.0141
6	-0.0038	-0.0070	-0.0097	-0.0127	-0.0116
7	-0.0003	-0.0092	-0.0112	-0.0133	-0.0119
8	0.0000	-0.0082	-0.0096	-0.0111	-0.0099
9	0.0010	-0.0067	-0.0076	-0.0086	-0.0078
10	0.0026	-0.0045	-0.0051	-0.0057	-0.0054
∞	0.0113	0.0069	0.0075	0.0069	0.0020

Table 5. Difference in recovery rates between full model and ones where all sequences after 5th, 4th, 3rd, 3nd and 1st are given same parameters (under the different write off policies WO(N) where N is number of times borrower stopped paying)

ACCEPTED MANUSCRIPT

$P(Pay_i^{j+1} \| Pay_i^j)$	**j (months into sequence)=1**	$j=2$	$j=3+$	**Chi-square $j=1$ vs $j=2$**	**Chi-square $j=2$ vs $j=3+$**
i (sequence number)=1	0.872	0.932	0.963	13.4	3.1
$i=2$	0.732	0.812	0.920	19.9	25.0
$i=3+$	0.593	0.715	0.901	166.3	206.8
$P(NoPay_i^{j+1} \| NoPay_i^j)$	$j=1$	$j=2$	$j=3+$		
$i=1$	0.409	0.777	0.962	761.9	71.3
$i=2$	0.413	0.735	0.958	420.5	73.8
$i=3+$	0.415	0.681	0.955	965.4	365.5

Table 6. Probabilities of transitions for hazard rate model obtained from case study data

i	1	2	3+	R
$RRM(i)$	0.0185	0.0185	0.02035	0.00185
$d(i)$	10	10	11	
D	541			

Table 7. *RRM(i)* average recovery rate per month for the hazard rate models from case study data with *d(i)* and *D* which give the discretized approximation

M	1	6	12	18	24	36	48
N=1	0.237%	1.201%	2.258%	4.357%	5.902%	8.555%	11.006%
N=2	0.237%	1.420%	3.338%	5.638%	8.136%	13.622%	17.866%
N=3	0.237%	1.972%	4.672%	7.849%	11.326%	18.557%	25.433%
N=∞	0.237%	1.972%	4.777%	8.397%	12.802%	23.706%	36.867%

Table 8. Recovery rates for different *WO(M,N)* policies using the hazard rate model where write-off occurs either after M periods into the collects process or at N^{th} time borrower stops paying

ACCEPTED MANUSCRIPT

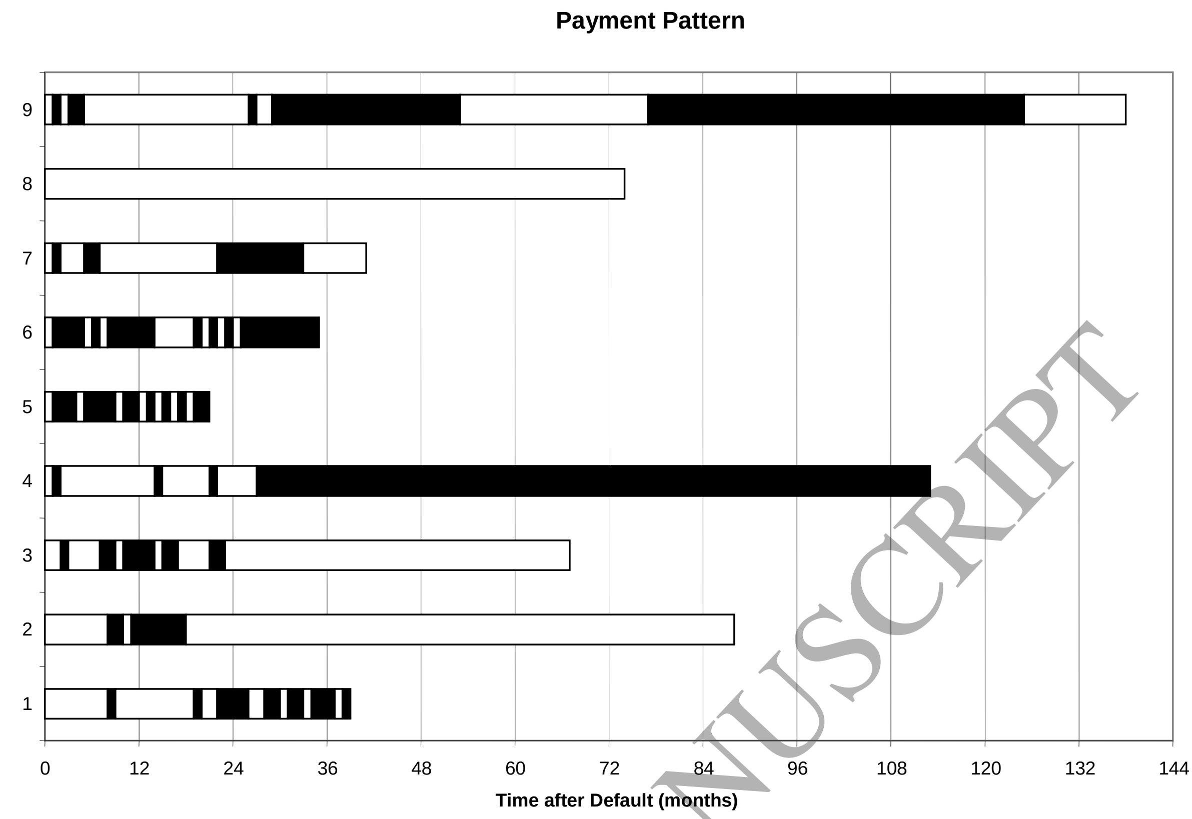

Figure 1. State space description of the payment sequences for 9 different defaulters (Note: Black when payment occurs; white when no payment)

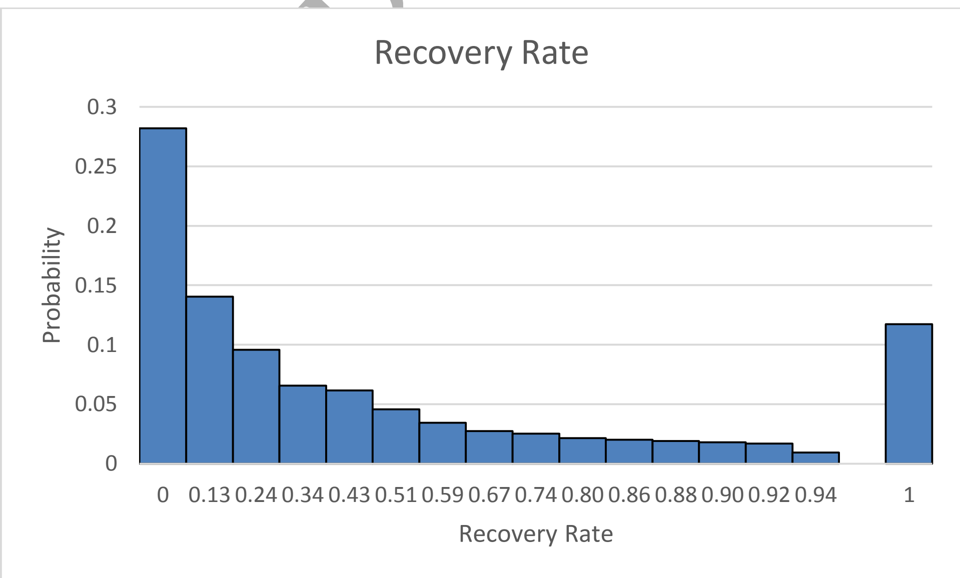

Figure 2: Recovery Rate (RR) distribution using full case study portfolio

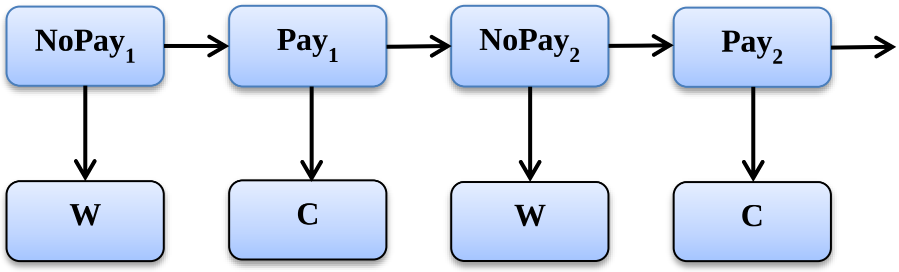

Figure 3. Transitions between states in repayment sequence model

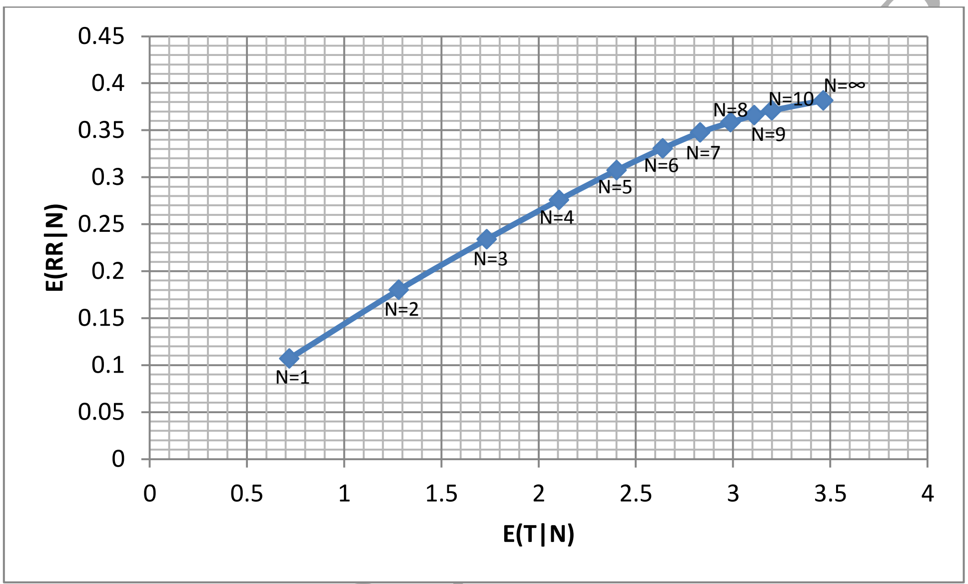

Figure 4. Trade-off between recovery rate and number of payment periods as the N in the write off policy *WO(N)* is 1,2,3,4,5,6,7,8,9,.10,∞ .

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

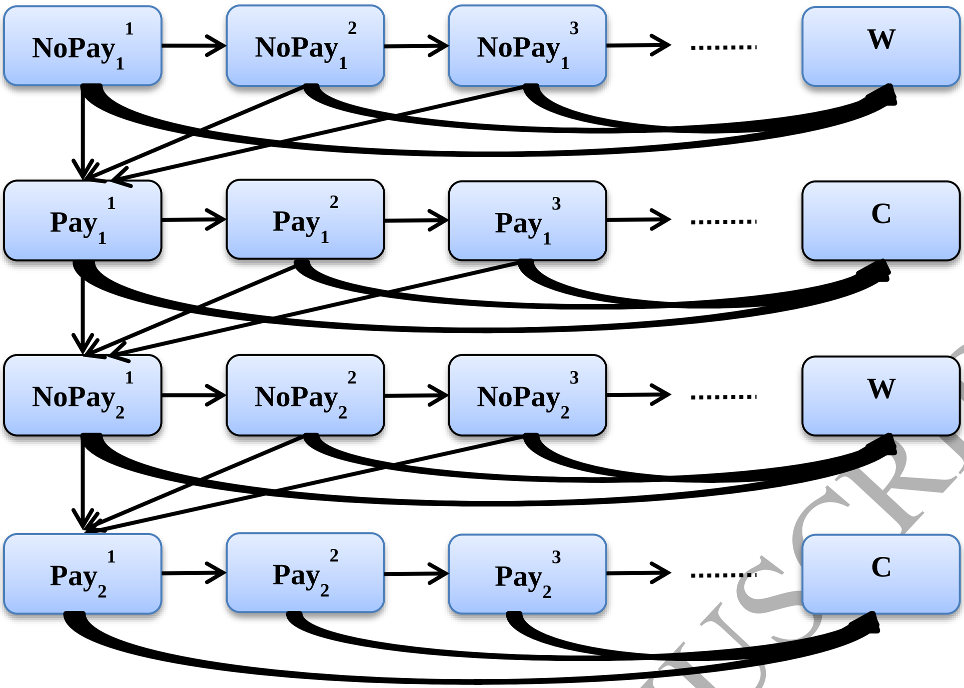

Figure 5.Transitions between states in hazard rate model

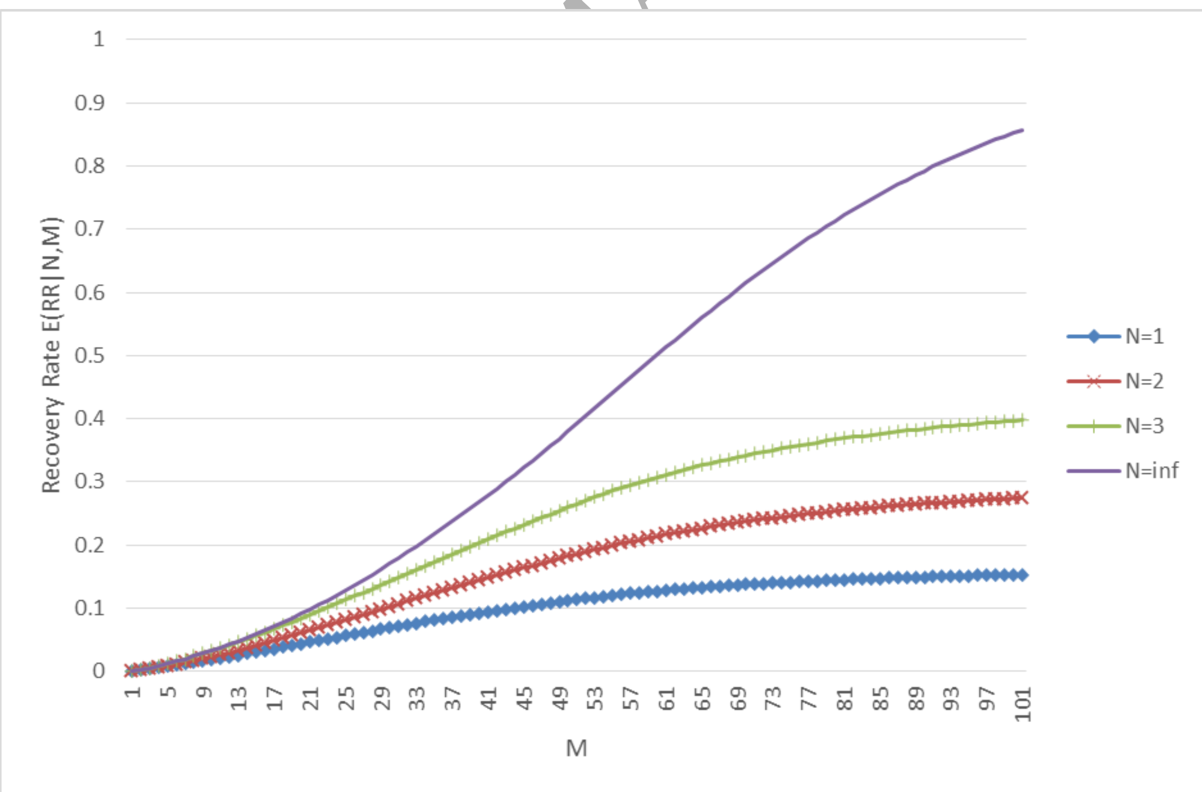

Figure 6. Recovery rates under different (*N*,*M*) write-off policies