Journal of Applied Mathematics & Bioinformatics, vol. x, no. x, 2011, x-x
ISSN: 1792-6602(print), 1792-6939(online)

Macro economic cycle effect on mortgage and personal loan default rates

Petrus Strydom[1]

Abstract

The aim of this paper is to apply a Gaussian process to decompose the time series of crude default rates into three components: age of the loan, quality of the loan and the exogenous economic environment. This is supported by the empirical result for a mortgage and personal loans portfolio based on five years of historic data. The Gaussian process does not impose an explicit parametric structure to the relationship between the three components and the default rate compared to other methodologies that assume a linear structure.

We find that the vintage and economic cycle components are more important drivers of the default rate compared to the age effect and this varies over the economic cycle. The contribution of the economic cycle component to the overall default (for both the mortgage and personal loan portfolios) range from between 20% to 50% depending the position in the economic cycle. In general the economic cycle effect is larger during economic stress.

JEL Classification :C11, C23, C53, E51, G21

Keywords: Credit Risk; Probability of Default; Forecasting; Gaussian process

[1] PhD Student, University of Witwatersrand. e-mail: pstrydom6@gmail.com,

Article Info: *Received* : January 29, 2017. *Published online* : xxxx

1 Introduction

The ability to forecast the credit risk outcome under various macro-economic scenarios is critical to support a wide range of applications such as; stress testing, scenario analysis, capital planning and an expected loss impairment assessment. The credit risk outcome is not only a function of the macro-economic environment. Loan specific characteristics such as the age of the loan or the credit quality also impacts the credit risk outcome. Isolating the effect of a change in the economic cycle on crude default rates requires a methodology to decompose the loan specific characteristics and economic cycle effects from historical trends.

Das and Stein 2008, Anderson *et al.* 2008 and Capozza and Van Order 2010 used time variant hazard rate models with an assumed linear structure to decompose the vintage and economic cycle effects. The models used to decompose the components and the granularity of the data varied, resulting in a range of estimates. In some studies the economic cycle was the most important component, where in other studies the vintage quality was the most important component. A key characteristic of the various decomposition models is the assumed logistic linear structure of the vintage and economic cycle components in the hazard rate specification. None of the studies included a comparison of the out of sample performance of the various models. This paper used a methodology set out by Zhang 2009 and Breeden 2007 to decompose the default rate for both a mortgage and personal loan portfolio of a South African bank. The model does not assume an explicit linear structure and captures a more complex non-linear relationship between the age-vintage and economic cycle components. This paper also includes an out-of-sample test of the Gaussian process model which was absent from other studies.

The model in this paper decompose the time series of crude default rates into three components: age of the loan, quality of the loan and the exogenous economic environment. This is termed maturity-exogenous-vintage (”MEV”) decomposition as discussed in Breenden 2007 and Zhang 2009. Different variations of linear model structures are regularly used in retail credit default rate models (see Canals-Cerda & Kerr 2015 and Malik & Thomas 2007).

The out-of-sample comparison is based on a forecasting model for the mortgage portfolio. This is used to test the out-of-sample performance of the MEV. The per account level performance data of the loan portfolio over a four year period from 2007 to 2010 is used for the MEV decomposition. We compare the actual default rate for the portfolio to the forecast from the MEV methodology on an out-of-sample basis (over 2013 and 2014). The MEV decomposition provides a very close fit to the actual default experience over the out-of-sample period.

2 Literature Study

Retail credit risk models focus on either the client level or portfolio level assessment of credit risk. Scoring models are used to assess the credit risk at a client level along with a corresponding mapping of a default rate to each credit score based on historic data. The portfolio view aggregates the client view to provide a portfolio level default rate. Thomas *et al.* 2005 provides a summary of the development in retail credit risk management, including an overview of the current issues. The original focus of credit risk modeling was to support the initial loan origination decision. This lead to developing various credit scoring approaches based on score cut offs. The development of the Basel I and Basel II regulatory requirements forced banks to shift the focus to estimating the probability of default and loss given default. Popular methods used by banks to calibrate the PD at a customer level are linear regression, logistic regression or mathematical programming methods. See Thomas *et al.*, 2002 for a discussion on credit scoring models and the application in PD calibrations. See Baesens 2003 for a comparison of 17 different type of methods and Altman & Saunders 1998 for a summary of existing methodologies.

The aggregated portfolio level view of credit risk is used to forecast the performance of the loan portfolio. Bellottie & Crook 2008, Bucay & Rosen 2001 and Rosh & Scheule 2004 developed various hazard rate or correlation focused methodologies where the default rate is linked to macro-economic variables. Malik & Thomas 2007 extended this by considering the behavioral score, age of the loan and macro-economic variables in the default rate estimate by either using a hazard rate or the Markov chain model.

Credit risk forecasting methodologies that ignore the credit score and age effect, and only consider the macro-economic variables will exclude important components in the forecasting methodology, see Breenden *et al.* 2007 for a discussion.

Decomposition methods

Das and Stein 2008 used 136 mortgage-backed transaction in the United States to determine the contribution of the economic cycle and vintage quality on a mortgage portfolio. Based on their methodology the vintage component is around double (200%) the economic cycle component during 2007-2008. Anderson *et al.* 2008 used aggregate portfolio level sub-prime mortgages in the United States to quantify the vintage quality and economic cycle components. The contribution of the two components is split 50:50 based on their empirical investigation. Capozza and Van Order 2010 also confirmed the 50:50 split between the economic cycle and vintage components based on aggregate foreclosure data. Capozza and Van Order 2010 applied the same model to a richer data set using loan level data, including the origination vintage. The economic cycle impact was the most important component of the default rate when controlling for the vintage components.

The data supporting these investigation as well as the methodologies applied to decompose the components are different. Das and Stein 2008 used a simulation based approach to estimate loan-level default rates where Anderson *et al.* 2008 used a time series of portfolio level default rates and a linear regression to regress a range of macro-economic and vintage quality proxies. The Das and Stein 2008 model requires a series of external models to calculate parameters such as prepayment and default rates based on the granular client level information such as credit scores and loan to value ratios. The Anderson *et al.* 2008 and Capozza and Van Order 2010 models are time variant hazard rate models with an assumed linear structure.

We use a Gaussian process to decompose historic default rates into an age, quality of loans at origination and exogenous or macro-economic cycle components. The Zhang 2009 methodology capture the heterogeneity of the portfolio into the explicit calibration of the vintage component. A Gaussian process was chosen instead of a linear structure to incorporate more complex non-linear relationships in the observed default rates.

Maturation- Exogenous- Vintage ("MEV") decomposition

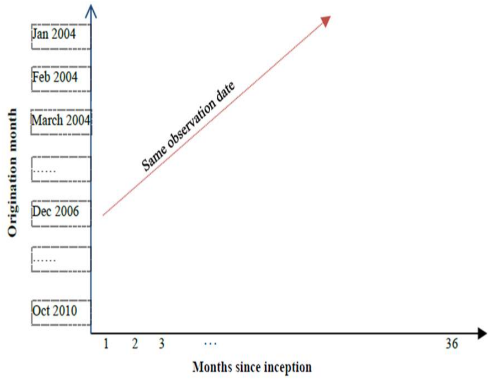
Zhang 2009 and Breenden 2007 developed a detailed methodology to decompose the default rate of various credit risk portfolios which share the duel time dynamic of origination, age of contract and the external environment. Zhang proposed a Gaussian process with an adoptive smoothing kernel function to decompose the observed credit risk into a Maturation- Exogenous- Vintage ("MEV") effect. This allows the practitioner to understand the observed credit experience in terms of time since origination (m), exogenous influenced by macro-economic conditions (t) and heterogeneity introduced from the various origination groups (v). The aim of this methodology is to develop a model to forecast the default rate for a particular origination group by considering only the three functions above. This paper compares the performance of the decomposition using the Gaussian process and a linear time variant hazard rate model.

3 Data

A vintage in the context of this paper comprises all loans originated in the same month. All loans from the same origination vintage have the same maturity or age development over the observation and projection period.

The origination month of the obliger is shown on the y-axis of figure 1 where the months since inception of the vintage is shown on the x-axis. The diagonal development represents the outcome under the same economic conditions, but across different origination vintages.

Figure 1: Overview of the format of input data



For example: loans written in Dec 2006 will be age 1 in Dec 2006, age 2 in Jan 2007 and ... age 36 in Nov 2009. The diagonal relates to the performance of the portfolio in the same period. For example: this will include loans originated in December 2006 at age 1, loans originated in November 2006 at age 2, ...

Vintage data as presented in this form is part of a class of panel data with a duel-time characteristic. The data shares the following cross sectional time series features:

- Origination month (v_j).
- Calendar observation month (t_{j+l}).
- And time since origination (m_l).

With J vintages and the credit performance of each vintage observed for L_j months. Where the Lj is different for each vintage. Each of the J vintages are observed at monthly intervals t_{j+l}, where $l = 1, 2, 3, ...L_j$.

In our case, each of the J vintages relate to a specific origination vintage v_j such that $v_1 < v_2 < ... < v_J$. The duration in force for each vintage is specified by m_l, this allows us to derived $t_{j+l} = v_j + m_l$.

Using the notation above the vintage data is presented in a panel format as

$$(m_l, t_{j+l}, v_j) \tag{1}$$

for $j = 1, 2, ...J, l = 1, 2, ..L_j$.

3.1 Mortgage data

An account level monthly snapshot of all mortgage accounts was provided from December 2006 to December 2011, this only included loans originated from January 2004 to December 2010. The performance status of the loans are tracked on a monthly basis. An account is classified as in default if the account missed 3 installments. Over 500 000 individual accounts were tracked over the observation period, with over 40 000 defaults identified. The crude default rate for origination month i at duration j is calculated as the number of defaults from origination month i between duration j and $j + 12$ divided by the number of accounts (excluding defaults) in origination month i at duration j.

The vintage data is denoted by $x_{j,l} = m_l, t_{j+l}, v_j$ with $j = 1, 2, ...J$, $l = 1, 2, ...L_j$. In this example j is the origination vintage from January 2004 (1) to December 2010 (49) and L_j refers to the number of monthly observations available for each origination vintage j. L_1 comprise of loans originated in January 2004. The observation period started in December 2006, thus the duration available for vintage L_1 was from 36 to 50 months. For the purpose of this model, L_j was capped at 36 months.

4 The MEV model

Observed default rates per the panel data

Each of the J origination vintages are observed for L_j months, this results in n observations, where $n = \sum^J \sum^{L_j} 1$. The crude default rate is observed for each of the n observations. Let $x_{j,l} \equiv (m_l, t_{j+l}, v_j)$ be the panel data of vintage v_j observed m_l months since origination, which will translate to calender month $t_{j+l} = v_j + m_l$. Let $y(m_l, t_{j+l}, v_j)$ be the observed default rate for the observation per the panel data $x_{j,l}$.

Let $\eta(x_{j,l})$ be a transformation function that maps the input data (m_l, t_{j+l}, v_j) to the default rate $y(m_l, t_{j+l}, v_j)$.

Formulation of $\eta(x_{j,l})$

The η function is broken down into a linear and non-linear portion

$$y(m_l, t_{j+l}, v_j) \approx \eta(x_{j,l}) = \mu(x_{j,l}) + Z(x_{j,l}) + \epsilon \tag{2}$$

where $\mu(x_{j,l})$ is a linear function of the panel data (included to capture the base or intercept impact or general linear trends), $\epsilon \approx N(0, \sigma^2)$ the random error, and $Z(x_{j,l})$ a zero-mean Gaussian process.

Definition 1 : Covariance structure of $Z(x_{j,l})$

Let $\mathbf{x}$ be a vector of the n panel observations $x_{j,l}$. Define the covariance between $Z(x_{j,1})$ -$\mu(x_{j,1})$ and $Z(x_{j,2})$ -$\mu(x_{j,2})$ as the $(j+1, j+2)^{th}$ element of the $n * n$ matrix K. K is a matrix of the covariance of each of the n inputs with each other. Let $\sigma^2 K(\mathbf{x}, \mathbf{x}')$ be the $n * n$ matrix, where $\mathbf{x}$' is the transpose of vector $\mathbf{x}$. The covariance of a specific panel data observation $x_{j,l}$ with the n observation is defined as $\sigma^2 K(\mathbf{x}, x_{j,l})$.

The Identification Lemma per Zhang 2009 requires us to break up the linear dependency of the panel data in the specification of the linear portion of η. This is required to ensure the model is identifiable. This is achieved by specifying the linear portion of η with reference to m and t only.

$$\mu(x_{j,l}) = \mu_0 + \mu_1 m_l + \mu_2 t_{j+l} \tag{3}$$

The linear portion of the model cannot include all three inputs per the panel data (m, t, v). We exclude v_j from equation 3 to satisfy this requirement. However, the formulation of $\mu(x_{j,l})$ can be simplified to $\mu(x_{j,l}) = \mu_0$ without any loss to the MEV decomposition. The linear portion of the model is very useful if there is a general trend in exogenous and maturity effect to be captured by a linear formulation.

Estimation of η

The solution requires a formulation of the expected default rate for each input data point $x_{j,l}$. This is achieved by estimating

$$E[\eta(x_{j,l})] = \mu(\hat{x_{j,l}}) + E[Z(x_{j,l}) + \epsilon|\tilde{z}] \tag{4}$$

where $\mu(\hat{x_{j,l}})$ is the linear estimation and $\tilde{z} = y(m_l, t_{j+l}, v_j) - \mu(\hat{x_{j,l}})$.

From Zhang 2009 and Rasmussen 2006 the expected value of the Gaussian process is

$$E[Z|\tilde{z}] = K(\mathbf{x}, x_{j,l})[K(\mathbf{x}, \mathbf{x}') + I]^{-1}(y - \hat{\mu}) \tag{5}$$

where $K(\mathbf{x}, x_{j,l})$ and $K(\mathbf{x}, \mathbf{x}')$ is per Definition 1.

The MEV decomposition

Zhang postulated $Z(x_{j,l})$ as a Gaussian process to allow the expected value to be defined in terms of the covariance functions, with an estimation based on the spline estimation technique. A Gaussian process was chosen as this requires less structure w.r.t the format of the function that maps the input data $(x_{j,l})$ to the observed default rates, but rather considers the distribution function of all possible functions that satisfy the mapping of the observed data in the training data. The GP estimation focuses on the replicating kernel, rather than specifies the structure of the mapping function.

Let $Z(x_{j,l}) = Z_f(m) + Z_g(t) + Z_h(v)$ be the sum of three independent Gaussian processes. Zhang assumed the three Gaussian processes are independent for simplicity, this model can be postulated to allow for the some dependency between the maturity, exogenous and vintage impacts. This requires a more complex structure, as it is difficult to ensure a positive definite cross-covariance function between the Gaussian processes (see Boyle & Frean 2005). The definition of $Z(x_{j,l})$ can be simplified by replacing one of the Gaussian processes (for example $Z_h(v)$) by a constant term or simple deterministic formula, then apply the same methodology to decompose the remaining two Gaussian processes and apply the back-fitting algorithm adjusted to estimate the deterministic portion.

The three Gaussian processes have the following practical explanations:

- $Z_f(m)$ represents the impact of the duration in force on the default rate.
- $Z_g(t)$ represents the exogenous influence of the macro-economic condition on the default rate.
- $Z_h(v)$ represents the vintage heterogeneity measuring the origination quality.

Let $\mathbf{m}$, $\mathbf{t}$ and $\mathbf{v}$ be a vector of the individual elements of the input panel data. For example: $\mathbf{v}$ is a vector representing the J vintages. Following the same covariance structure as per Definition 1, let $\sigma_f^2 K_h(\mathbf{v}, \mathbf{v}')$ be a $J * J$ matrix of the covariance of each of the J vintages with each other. The same definition apply for $\sigma_f^2 K_h(\mathbf{m}, \mathbf{m}')$ and $\sigma_g^2 K_h(\mathbf{t}, \mathbf{t}')$.

By the independence assumption

$$K(x_{j,l}, x_{j,l+1}) = \frac{\sigma_f^2}{\sigma^2} K_f(m_l, m_{l+1}) + \frac{\sigma_g^2}{\sigma^2} K_g(t_{j+l}, t_{j+l+1}) + \frac{\sigma_h^2}{\sigma^2} K_h(v_j, v_j) \qquad (6)$$

with $\lambda_f = \frac{\sigma_f^2}{\sigma^2}$,$\lambda_g = \frac{\sigma_g^2}{\sigma^2}$ and $\lambda_v = \frac{\sigma_h^2}{\sigma^2}$.

From proposition 3.2 and 3.3 in Zhang 2009 the additive separability properties of the spline estimator can be expressed as

$$\eta(x_{j,l}) = \mu(x_{j,l}) + \sum_{i=1}^{L} \alpha_i K_f(m_l, m_i) + \sum_{i=1}^{J+L} \beta_l K_g(t_{j+l}, t_i) + \sum_{i=1}^{J} \gamma_j K_h(v_j, v_i) \quad (7)$$

with coefficients determined by,

$$\min[||y - \mu(\hat{x}_{j,l}) - \tilde{\Sigma}_f \alpha - \tilde{\Sigma}_g \beta - \tilde{\Sigma}_h \gamma||^2 + \lambda_f ||\alpha||_{\Sigma f}^2 + \lambda_g ||\beta||_{\Sigma g}^2 + \lambda_h ||\gamma||_{\Sigma h}^2]$$

where $\tilde{\Sigma}_f = K_f(\mathbf{m}, m_i)_{n*1}$ and $\Sigma_f = K_f(m_i, m_i)_{1*1}$.

Definition 2 : The spline solution

The MEV decomposition defines the Gaussian process formulation per equation 2 as the sum of three Gaussian processes. The expected value of the Gaussian process is defined per equation 5. A spline formulation of equation 5 is used as an estimation technique. The spline solution is used to define equation 7 as the spline estimator of the expected value of the Gaussian process. Equation 7 formulates the expected default rate for each input from the panel data $x_{j,l}$. The isolated impact for panel data $x_{j,l}$ of the duration in force (m_i), exogenous influence (t_{j+l}) and the vintage impact (v_j) is shown in equation 7 as $\sum_{i=1}^{L} \alpha_i K_f(m_l, m_i)$, $\sum_{i=1}^{J+L} \beta_l K_g(t_{j+l}, t_i)$ and $\sum_{i=1}^{J} \gamma_j K_h(v_j, v_i)$.

Evaluating the three components for all the input data ($j \in J$ and $l \in L_j$) allows a graphical representation of the duration, exogenous influence and vintage components. Based on the above, this is a graphical representation of the expected value of the Gaussian process, not the actual process. This graphical representation will be termed $Z'_f(m)$, $Z'_g(t)$ and $Z'h(v)$. See figures 2, 3 and 4.

Zhang 2009 proposed a back-fitting procedure to estimate the above parameters (see appendix 1).

The work by Zhang 2009 established the link between the Gaussian process $Z_g(t)$ and the adoptive smoothing spline function. Where $E[Z_g(t_{j+l})] = \sum_{i=1}^{j+l} \beta_i K_g(t_{j+l}, t_i)$. This allows us to estimate the $Z_g(t_{j+l})$ for each of the n panel data inputs. The same formulation holds to estimate $E[Z_h(v_j)]$ and $E[Z_f(m_l)]$.

5 Results: MEV decomposition

The MEV methodology was fitted to a retail mortgage loan portfolio. The structure of the kernel functions using the back-fitting algorithm is shown below.

- $K_f(m)$ - Maturity curve, based on exponential kernel with $k = 1$. With $m = 1$ to 36.
- $K_g(t)$ - Exogenous influence curve, based on Mater-family kernel with $k = 2/3$. With $t = 1$ to 49.
- $K_h(v)$ - Origination vintage curve, based on Mater-family kernel with $k = 2/3$. With $v = 1$ to 72.

The θ_g, θ_f and θ_h parameters required to calibrate the covariance functions K_g, K_f and K_h per equation 7 are shown in table 1.

Table 1: MEV parameter calibration

Parameter	value
μ_0	0.02605
μ_1	0.02605
μ_2	0.02605
θ_g	2.6
θ_f	6
θ_h	3.5

The parameters α, β and γ from equation 7 are vectors estimated in step one and two and are the spline formulation of the Kriging. Per definition 2, the spline formulation provides an estimate of the expected value of the Gaussian processes, with a graphical representation per $Z'_f(m)$, $Z'_g(t)$ and $Z'_h(v)$. Figure 2 shows a graphical representation of the age impact on the observed default rates. A positive value is an add-on to the estimated default rate and thus indicates a deterioration in the credit quality of the portfolio. Based on this, the credit risk of a mortgage increases over the first 12 months from origination, where after this risk decreases. The contribution of the age effect on the default rate is negative from duration 27 onwards.

Figure 2: Mortgages: Age impact $Z'_f(m)$

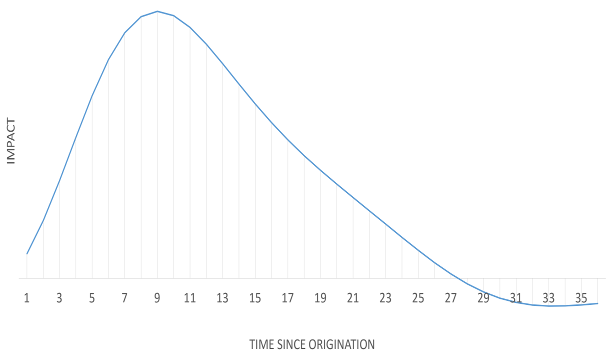


Figure 3 shows a graphical representation of the exogenous impact on the observed default rates. The figure shows large positive values during the economic crisis from April 2008 to June 2009, indicating the deterioration of the credit quality of all loans during this period. The exogenous impact on credit risk subsequently improved for six months from June 2009 to December 2009 with a subsequent deterioration during 2010.

Figure 3: Mortgages: Economic cycle impact $Z'_g(t)$

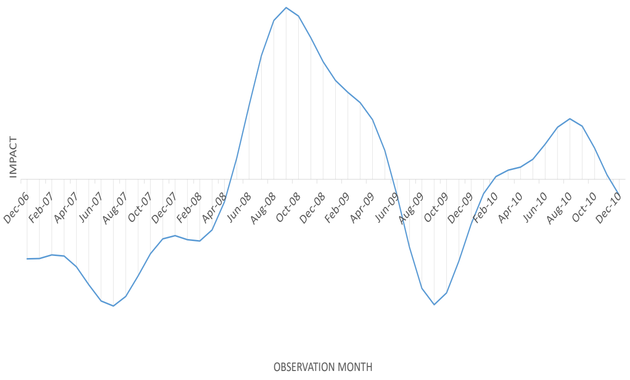
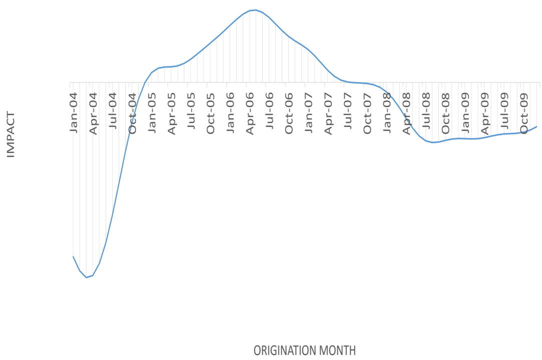


Figure 4 shows a graphical representation of the vintage impact on the observed default rates. The figure shows the lower credit quality of loans originated between January 2005 and June 2007. The lower quality of loans originated prior to 2008 and 2009, together with the economic cycle effect, explains the large increase in the observed default rates during this period of stress. The sudden changes in the vintage components related to policy decisions made by the bank, such as the loan-to-value requirements for new loans.

Figure 4: Mortgages: Vintage impact $Z_h'(v)$



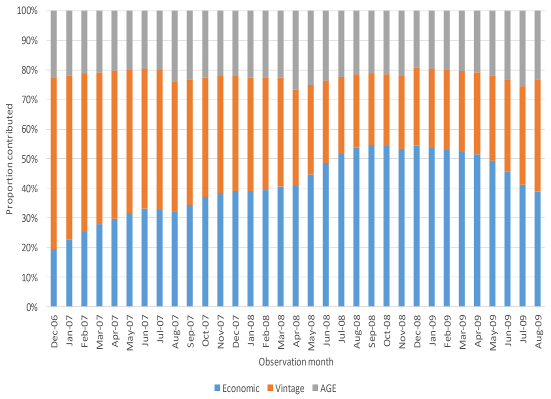
The graphical representation of the impact of the age, exogenous and vintage components on the observed default rate provides a transparent and intuitive view for stakeholders.

$Z_g'(t)$ represents the impact on the crude default rate due to the changes in the macro-economic environment, which will facilitate the projection of the default rate under various macro-economic scenarios. This is achieved by relating this function to macro-economic drivers via a regression method.

6 Results : Decomposition impacts

The sum of $Z_f(m)$, $Z_g(t)$ and $Z_h(v)$ provides the default rate per vintage. This can be combined across all the historical vintages over the observation period to calculate the portfolio level default rate. This allows us to calculate the cumulative impact of each of these components on the default rate over the historic time period. Figure 5 show the contribution of the age, economic and vintage components on the portfolio level default rate for Mortgages over the observation period (2006 to 2009) per the MEV methodology. The age impact is fairly consistent at 20% of the observed default rate over the period. The vintage impact range from 24% - 58% of the observed default rate, with the economic cycle impact between 19%-55%.

Figure 5: Portion of default rate explained by the MEV components



The economic cycle impact explained around 19% of the portfolio default rate at the start of the observation period (December 2006). This impact increased as the South African macro-economic cycle deteriorated. The mortgage portfolio default rate peaked in October 2008, this coincide with the period when the economic cycle explained up to 55% of the portfolio level default rate, up from 19% at the start of the period. This also coincide with the top of the interest rate cycle in South Africa. Note the economic cycle contribution to the portfolio level default rate dropped back to 39% by August 2009. The prime lending rate decreased by more than 200 basis points over this time period and the South African macro-economic cycle improved as measured by the gross domestic product.

The vintage impact on the mortgage portfolio default rate start at around 58% at the start of the observation period. The vintage impacts reduce to 24% in October 2008 before increasing back to 38% in August 2009. This empirical analysis confirms that the vintage and economic cycles impacts are the most important factors impacting the default rate of mortgage portfolios in South Africa. However the impact of both these components on the default rate varies as the macro-economic cycle changes. The vintage impact is the most important driver during benign macro-economic conditions.

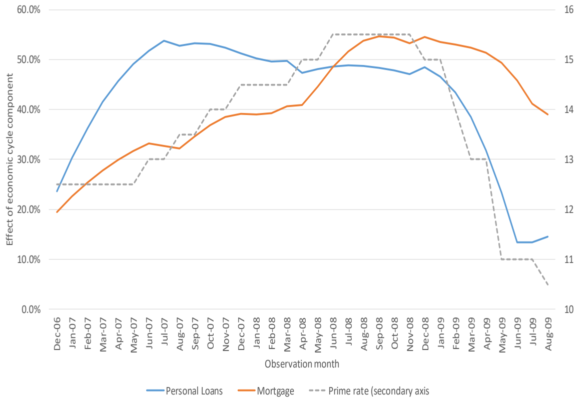
The same MEV decomposition methodology was followed for a personal loans portfolio. The age component also contribute around 20% for the personal loans portfolio increasing to 30% towards 2009 due to the significant growth in unsecured lending in South Africa, effectively reducing the average age of loans. Table 2 compares the maximum and minimum contribution of the macro-economic cycle and vintage components on the default rate over the same calibration period (December 2006 to August 2009). The overall effect of the quality of the vintage and the economic cycle effect is similar between personal loans and mortgages. The large increase in the age effect in 2009 explain the difference in minimum contribution of the economic cycle between personal loans and mortgages (the 13.4% versus the 19.5%).

Table 2: Contribution of vintage and macro-economic effect on default rates

Component	Maximum	Minimum
Mortgage : Vintage	24.1%	57.8%
Personal Loans : Vintage	26.2%	56.1%
Mortgage : Economic cycle	19.5%	54.7%
Personal Loans : Economic cycle	13.4%	53.8%

The impact of the vintage quality varies over the business cycle as the bank varies its loan origination criteria and risk appetite. Similarly the macro- economic effect varies as the economic cycle changes. Figure 6 compares the contribution of the macro-economic effect to the overall default rate for the personal loans and mortgage portfolios over the observation period. The South African prime lending rate is shown on the secondary axis as an indicator of the economic cycle. Although the contribution of the macro-economic effect on the default rate is similar between the two portfolios the timing of this effect differs. The effect of the macro-economic conditions seems to impact the personal loan portfolio prior to the mortgage portfolio. The economic effect on the mortgage portfolio seems to be largely interest rate driven where the personal loan portfolio is more sensitive to other lead indicators of changes in the interest cycle. The sharp recovery from 2009 onwards impacted the macro-economic components of both the portfolios at the same time.

Figure 6: Portion of default rate explained by macro-economic effect



6.1 Out-of-sample methodology

The MEV decomposition was calibrated based on defaults observed over 49 months (December 2006 to December 2010). We obtained the portfolio level default rate of the mortgage portfolio for 2013 and 2014. The out-of-sample comparison require us to use the MEV decomposition calibrated over December 2006 - December 2010 to estimate the portfolio level default rate for 2013 and 2014 to compare against the actual default rate observed.

Per the MEV methodology and following equation 2, the forecast of the default rate per vintage is the sum of the linear portion ($\mu(x_{j,l})$) and non-linear portion ($Z_f(m) + Z_g(t) + Z_h(v)$). No forecast of the vintage or age components is required, except for new vintages. The default rate forecast requires a projection of the exogenous impact ($Z_g(t)$) where t expands beyond the calibrated period (December 2010). A simple linear regression model is used estimate the macro-economic factors that explain $Z'_g(t)$. The actual outcome of these macro-economic factors over 2013 and 2014 is used to estimate $Z'_g(t)$ for 2013 and 2014. This allows us to calculate the out-of-sample performance of the MEV model.

Havrylchyk 2010 developed a regression type model to empirically test the impact on the credit loss due to a change in a set of macro-economic variables in the South African market. We use the MEV decomposition discussed to first isolate the exogenous impact ($Z'_g(t)$) from the observed default rate time series, before applying a similar approach as Havrylchyk 2010 to develop a regression model with a series of macro-economic inputs as independent variables to estimate $Z'_g(t)$.

The linear regression model has the following form:

$$Z'_g(t) = Factor_1 * \beta_1 + Factor_2 * \beta_2 + Factor_3 * \beta_3 + Factor_4 * \beta_4 + \varepsilon. \quad (8)$$

A linear regression method is used to estimate $\beta_1, \beta_2, \beta_3$ and β_4. Equation 8 is used to estimate the exogenous component based for a given set of macro-economic variables for Factors 1-4. The exogenous component is combined with the age and vintage components to forecast the portfolio level default rates.

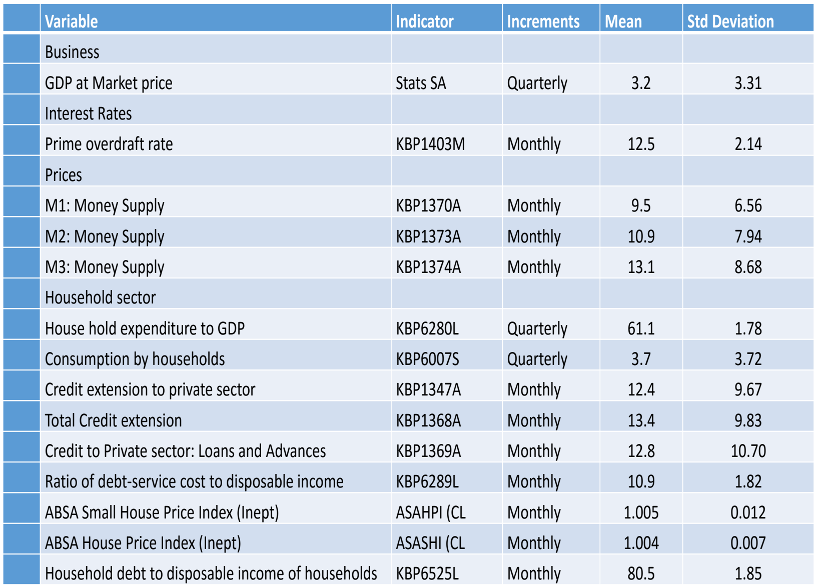
Independent variables considered for regression model

Macro-economic variables based on the following categories were considered:

- Business
- Interest rates
- Price Stability
- Household

The historic time series of economic variables were obtained from the South African Reserve Bank ("SARB") website. Figure 7 shows the factors and basic statistics over the calibration period.

Figure 7: Macro-economic variables considered



Variable	Indicator	Increments	Mean	Std Deviation
Business				
GDP at Market price	Stats SA	Quarterly	3.2	3.31
Interest Rates				
Prime overdraft rate	KBP1403M	Monthly	12.5	2.14
Prices				
M1: Money Supply	KBP1370A	Monthly	9.5	6.56
M2: Money Supply	KBP1373A	Monthly	10.9	7.94
M3: Money Supply	KBP1374A	Monthly	13.1	8.68
Household sector				
House hold expenditure to GDP	KBP6280L	Quarterly	61.1	1.78
Consumption by households	KBP6007S	Quarterly	3.7	3.72
Credit extension to private sector	KBP1347A	Monthly	12.4	9.67
Total Credit extension	KBP1368A	Monthly	13.4	9.83
Credit to Private sector: Loans and Advances	KBP1369A	Monthly	12.8	10.70
Ratio of debt-service cost to disposable income	KBP6289L	Monthly	10.9	1.82
ABSA Small House Price Index (Inept)	ASAHPI (CL	Monthly	1.005	0.012
ABSA House Price Index (Inept)	ASASHI (CL	Monthly	1.004	0.007
Household debt to disposable income of households	KBP6525L	Monthly	80.5	1.85

Multivariate analysis:

A stepwise regression methodology was used to select a linear regression model. A cluster and correlation analysis was performed to identify high correlations. The cluster analysis identified three main clusters, and confirmed that a number of the macro-economic variables are highly correlated. Only one variable per cluster was selected for the final regression model. The final regression model has an R^2 of 84% over the calibration period.

Table 3 show the final coefficients and corresponding P-values supporting the stepwise regression.

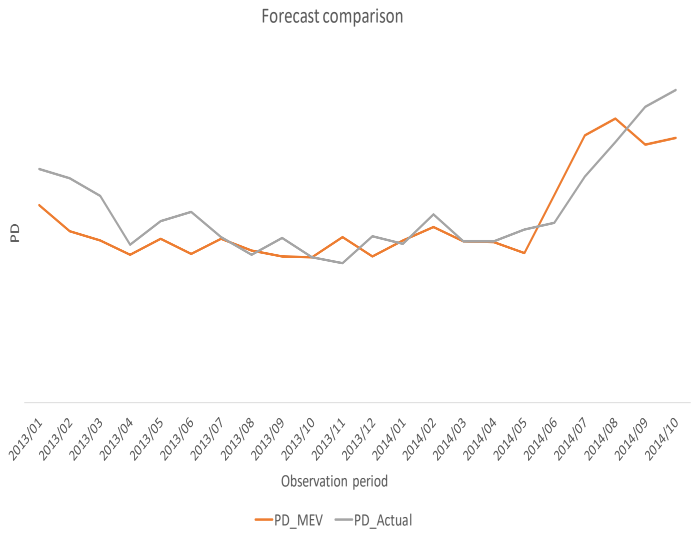
Figure 8 shows the default rate forecast for the mortgage portfolio over 2013 and 2014, using the results from the MEV methodology previously described(PD_{MEV}). The results are compared to the actual observed default rates over the same period (PD_{Actual}). The estimated default rates using the MEV decomposition for 2013 and 2014 closely follow the the actual default rate. This period was characterised by a 50 basis point drop in the prime lending rate in August

Table 3: Regression coefficients for $Z'_g(t)$

Variable	Coefficient	P-Value
Intercept	0.15531	$< .0001$
Prime	0.00092624	$< .0001$
M2 supply	-0.00020266	$< .0001$
House price index	-0.13976	$< .0001$
Household debt to income	-0.00030455	0.0038

2012, with a corresponding 50 basis point increase in January 2014 and 25 basis points in July 2014. The out-of-sample model based on the MEV decomposition accurately captures the impact change in the base lending rate, as shown by the strong out-of-sample test.

Figure 8: Default rate forecast using the MEV and simple regression methods



7 Acknowledgments

I would like to thank Dr D, Wilcox for her helpful comments on this paper.

8 Appendix

8.1 Back-fitting algorithm

The dependency of the panel data results in a collinearity problem in estimating the individual Gaussian processes. The back-fitting procedure approach follows a step-by-step approach where each of the Gaussian processes is estimated in isolation. This better deals with the collinearity problem than a maximum likelihood estimation.

Assume a pre-specified structure for the three covariance functions $K_f(\mathbf{m}, \mathbf{m}')$, $K_g(\mathbf{t}, \mathbf{t}')$ and $K_h(\mathbf{v}, \mathbf{v}')$. Assume either an exponential of Matern function completely described by θ_f, θ_g and θ_v. Matern is a class of covariance functions $u(k, \theta, r)$ with parameters k, $p = k + 0.5$, θ and with $r = x_i - x_j$ the elements in the covariance matrix. See Guttorp and Gneiting 2006 for a discussion on Matern functions. The Matern function is defined as:

$$u(k, \theta, r) = \exp(-\frac{\sqrt{2kr}}{\theta})(\frac{\Gamma(p+1)}{\Gamma(2p+1)})\sum_{i=1}^{p}\frac{(p+1)!}{(p-1)!}(\frac{\sqrt{8kr}}{\theta})^{p-i}. \tag{9}$$

The covariance functions K_f, K_g and K_h are defined by θ_f, θ_g and θ_v. An optimal solution may be found by setting the partial derivative of equation 7 to zero to solve the remaining parameters. However, this requires a process to first estimate θ and λ. The back-fitting algorithm overcomes this problem.

Cross-validation is a methodology used in spares data statistics where one input dataset is used to both fit and test a specific model. A leave-one-out method trains the function on both the complete input dataset as well as on the dataset less one data point. The cross-validation score is measured as the difference in the actual versus expected, using the various leave-one-out fits. Generalised cross-validation (”GCV”) is a well-established method used to fit smoothing splines, as the smoothing matrix supporting the spline has a direct link to the GCV measure, see Gelub *et al.* 1979.

The back-fitting algorithm is based on the principle of GCV as described in Bates *et al.*, 1986 and Gelub *et al.*, 1979. A ridge regression methodology is used to apply the GCV methodology to determine the parameters θ,λ_f, λ_g and λ_v.

Define the complete set of parameters by $[\mu, \alpha, \theta_f, \lambda_f, \beta, \theta_g, \lambda_g, \gamma, \theta_h, \lambda_h, \sigma^2]$. Where λ are the smoothing parameters under the adoptive smoothing spline structure; θ_i the parameters specifying the covariance function; α, β, γ the reproducing kernel function and μ, σ the mean and variance.

The six steps in the back-fitting algorithm are:

Step 1: Set the initial value of μ by the ordinary least squares.

Step 2: Create pseudo data $\tilde{y} = y - \hat{\mu(x)} - \Sigma_g\beta - \Sigma_h\gamma$. Estimate α by ridge regression and determine (θ_f, λ_f) by minimising the GCV(θ_f, λ_f).

Step 3: Create pseudo data $\tilde{y} = y - \hat{\mu(x)} - \Sigma_f\alpha - \Sigma_h\gamma$. Estimate β by ridge regression and determine (θ_g, λ_g) by minimising the GCV(θ_g, λ_g).

Step 4: Create pseudo data $\tilde{y} = y - \hat{\mu(x)} - \Sigma_g\beta - \Sigma_f\alpha$. Estimate γ by ridge regression and determine (θ_v, λ_v) by minimising the GCV(θ_v, λ_v).

Step 5: Re-estimate μ for $Z(x)$ given $[\theta_f, \lambda_f, \theta_g, \lambda_g, \theta_v, \lambda_v]$.

Step 6: Repeat steps 2-5 until convergence. Once convergence is reached, obtain σ as derived from the multivariate setup for Z(x).

This methodology is set out in Zhang 2009.

References

[1] Altman EI & Saunders A, 1998, *Credit risk measurement: Developments over the last 20 years*, Journal of Banking & Finance, **21**, pp. 1721-1742.

[2] Anderson CD % Capozza DR & Van Order R, 2008 *Deconstructing the Subprime Debacle Using New Indices of Underwriting Quality and Economic Conditions: A First Look*, Available from http://ssrn.com/abstract=116809080739.

[3] Baesens B, 2003, *Developing intelligent systems for credit scoring using machine learning techniques*, PhD Thesis, Katholieke Universitiet Leuven, Belgium.

[4] BANK OF INTERNATIONAL SETTLEMENTS, 2006, *International convergence of capital measurement and capital standards*, Basel Committee on Banking Supervision.

[5] Bates DM & Lindstorm MJ & Wahba G & Yandell BS, 1986, *GCVPACK - Routines for Generalized Cross Validation*, University of Wisconsin, WI.

[6] Bellotti T & Crook J, 2009, *Credit scoring with macroeconomic variables using survival analysis*, Journal of Operational Research Society, **60**, pp. 1699–1707.

[7] Boyle P & Frean M, 2005, *Dependent Gaussian Processes*, Advances in Neural Information Processing Systems, **17**, pp. 217–224.

[8] Breenden JL, 2007, *Modeling data with multiple time dimensions*, Computational Statistics & Data Analysis, **51**, pp. 4761–4785.

[9] Breenden JL & Thomas L, 2008, *The relationship between default and economic cycles for retail portfolios across countries*, Journal of Risk Model Validation, **2**, pp. 11–47.

[10] Breenden JL & Thomas L & Mcdonald J, 2007, *Stress Testing Retail Loan Portfolios with Duel-time Dynamics*, The Journal of Risk Model Validation, **2**, pp. 43–62.

[11] Bucay N & Rosen D, 2001, *Applying Portfolio Credit Risk Models to Retail Portfolios*, Journal of Risk Finance, **2(3)**, pp. 35–61.

[12] Canals-Cerda JJ & Kerr S, 2015 *Forecasting credit card portfolio losses in the Great Recession: a study in model risk*, Journal of Credit Risk, **11**, pp. 29–57.

[13] Capozza DR & Van Order R, 2010 *The Great Surge in Mortgage Defaults 2006-2009: The Comparative Roles of Economic Conditions, Underwriting and Moral Hazard*,Journal of Housing Economics, **20**, pp. 141–151.

[14] Das A & Stein RM, 2009 *Underwriting versus economy: a new approach to decomposing mortgage losses*, The Journal of Credit Risk, **5**, pp, 19-41.

[15] Guttorp P & Gneiting T, 2006, *Studies in history of probability and statistics on the Matern correlation family*, Biometrika, **93**, pp. 989–995.

[16] Gelub GH & Heath M & Wahba G, 1979, *Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter*, Technometrics, **21**.

[17] Havrylchyk O, 2010, *A macroeconomic credit risk model for stress testing the South African banking sector*, South African Reserve Bank, **Working paper 3**.

[18] Malik M & Thomas L, 2007, *Modeling Credit Risk of Portfolio of Consumer Loans*, School of Management, University of Southampton, United Kingdom, Available from http://ssrn.com/abstract=1287845.

[19] Rasmussen CE, 2006, *Gaussian Process for Machine Learning*, MIT press.

[20] Rosch D & Sscheule H, 2004, *Forecasting Retail Portfolio Credit Risk*, Journal of Risk Finance, **5(2)**, pp. 16–32.

[21] Thomas LC & Edelman DB & Crook JN, 2002, *Credit Scoring and its Applications*, Paper presented at the 50th anniversary of the Society for Industrial and Applied Mathematics, Philadelphia.

[22] Thomas LC & Oliver RW & Hand DJ, 2005, *A survey of the issues in consumer credit modeling research*, Operational Research Society, **56**, pp. 1006–1015.

[23] Zhang A, 2009, *Statistical Methods in Credit Risk Modeling*, PhD Thesis, The University of Michigan, Mich, Available from http://deepblue.lib.umich.edu/bitstream/2027.42/63707/1/ajzhang_1.