Agricultural Economics Society

97th Annual Conference 2023

Full Programme

The University of Warwick

Check out our Twitter account @AgEconSoc. Feel free to Tweet your experiences at the Conference using #AES_2023.

A typology of Malian farmers and their credit repayment performance – An unsupervised machine learning approach

Tim Ölkers, Shuang Liu, Oliver Mußhoff [*]

March 1, 2023

Abstract

The availability of formal credit is crucial for the development of the agricultural sector as it can enhance farmers' purchasing power to acquire inputs and agricultural technology. This, in turn, can increase productivity and resilience throughout the sector. Therefore, the analysis of bank client and loan data in the agricultural sector in a developing country is of interest. We explore the question of who the clients of agricultural credit are and whether they can be clustered into different groups by using an unsupervised machine learning technique. We also investigate whether the loan repayment performance of these clusters differs based on various logit regressions. According to our results, there are 3 different clusters of farmers in Mali that differ by personal characteristics (such as age or gender) as well as credit demand characteristics (e.g., loan amount, interest rates, credit duration, number of credits). Each cluster that differs in their characteristics demonstrates a dissimilar repayment performance. Hence, different instruments as well as communication designs are needed to meet the financial needs of the different clusters and to strengthen the resilience of different groups of farmers in Mali. Our findings provide an important foundation for the design of future agricultural policies and financial products for the agricultural sector as they emphasise the heterogeneity of agricultural lenders in general.

Keywords: Agricultural Credit, Agricultural Policy, Loan Demand, Mali, Machine learning, Farmer Typology

JEL Code: C55 (Large Data Sets), G20 (Financial Institutions), G21 (Banks, Depository Institutions, Micro Finance Institution), O13 (Agriculture), O16 (Financial Markets, Saving and Capital Investment, Corporate Finance and Governance), Q14 (Agricultural Finance), Q18 (Agricultural Policy; Food Policy)

[*] Department of Agricultural Economics and Rural Development, University of Göttingen, Germany

1 Introduction

Climate change has several implications for agricultural production (Carter et al., 2018; Cui & Xie, 2022; Dumortier et al., 2021; Faye et al., 2023; Fujimori et al., 2022; Webber et al., 2018). Hertel & Rosch (2010) show how agriculture is a major pathway for transferring the impacts of climate change to the poor. It is expected that extreme weather events such as erratic rainfall patterns and warmer environments increase (IPCC, 2022). This makes the discussion of appropriate and precise agricultural policies to support adaptation to climate change particularly interesting. Additional liquidity is needed to increases the resilience of the agricultural sector through adoption strategies (Batung et al., 2023; Flory, 2018; Suri & Udry, 2022).

Many farmers and businesses in general in the Gloabl South are credit constraint (Banerjee & Duflo, 2014). However, better access to formal credit, including microcredit, commercial and agricultural banks, can help rural households mitigate risks by improving the access to inputs and other technology to modernize agriculture, thereby increasing resilience as well as profits (Fafchamps et al., 2014; Khandker & Koolwal, 2016). Lack of access to sufficient financial services is considered a hindrance to the economic development of the agricultural sector. Nevertheless, even when financial services are accessible, high default rates are one of the main problems of financial institutions in the Global South in general. This problem also harms the sustainability of these financial institutions (Weber & Musshoff, 2012, 2017). Various factors affect the repayment behaviour of credit clients. Agricultural income is relatively volatile due to its dependence on factors such as weather and pests, market risks and input prices, especially for smallholder farmers in developing countries (Rosenzweig, 2001; Stephens & Barrett, 2011). Additionally, food price seasonality affect the available funds to repay credit (Gilbert et al., 2017). The business cycles in the agricultural sector, with their different phases of crop cycles, also determine the availability of liquidity for farmers (Channa et al., 2022). All these mentioned points that affect income levels can impact the proportion of funds that can be allocated towards repaying loans for agricultural borrowers, leading to an increased likelihood of defaulting on their loans (Barry, 2001). According to the literature, further mentioned determinants that affect non-performing loans are attributable to issues like moral hazard (Banerjee, 2013; Hellmann et al., 2000), insufficient monitoring (Ghosh et al., 2020), and credit characteristics (Godquin, 2004; Weber & Musshoff, 2017). However, Raghunathan et al. (2011) show that agriculture loans have a strong positive influence on repayment efficiency of borrowing groups in India.

Furthermore, agricultural farm structures all over the world, including the Global South, have been and continue to be in flux. The changing farm structure of this complex entity also shapes the discussion about the appropriate design of agricultural policy, because each type of farm has different policy needs. This raises the question of who the farmers are, how they can be characterized, and whether these farmers can be grouped based on a broad set of characteristics that includes personal characteristics such as age, gender and marital status as well as information on their credit history (e.g., average duration and amount of credit). A comprehensive understanding of the structure of farmers is a crucial basis for the development and implementation of accurate, target group-oriented policy measures. Farmers and farm structures are predominantly classified on the basis of only one dimension (e.g., farm size). However, this only leads to an under complex understanding and does not reflect the multiplex reality of farmers in the Global South. Hence, a multidimensional perspective, encompassing divers factors to describe the farmer and their business, is needed as a basis for appropriate policy design. On the one hand, several typologies for farmers already exist. On the other hand, many of those typologies focus on farmers in a certain country of the Global North (Bartkowski et al., 2022; Graskemper et al., 2021), only on certain types of farmers e.g., dairy farmers (Methorst et al., 2017), small farms (Guarín et al., 2020), or pick different geographical regions within a country instead of focusing on a whole country (Daloğlu et al., 2014). It is therefore difficult to generalize these farm cluster results. Hence, the value for policy recommendations is often limited to a particular context.

To the best of our knowledge, there exist no typology of farmers based on farmers characteristics as well as their credit behaviour, hence, combining client and credit uptake data. Therefore, the objective of this study is to empirically analyse (1) whether there are different clusters of agricultural farmers and how are these clusters characterized and (2) whether these heterogeneous farmer clusters differ in their repayment behaviour. In our empirical work, we rely on a large data set of a commercial bank in Mali, consisting of around 10,000 granted credits for Malian farmers. The data set covers loans issued between 2010 and 2022 and includes borrower characteristics, credit characteristics, and repayment results. In the first step, we follow the estimation strategy of Graskemper et al. (2021) and provide a depiction of Malian farmers based on the sociodemographic characteristics as well as credit uptake data using unsupervised machine learning (ML) clustering analysis (Partitioning Around Medoids (PAM)). We then use the Elbow method to determine optimal number of clusters. Through this method we can correct for a potential researchers' bias towards specific topics by using an unsupervised ML approach (Graskemper et al., 2021). In the second step, we follow the estimation strategy by Weber & Musshoff (2012), De Andrés et al. (2021), and Grohmann et al. (2020) to examine the loan repayment for each cluster of all agricultural loans disbursed by the commercial bank.

Mali is one of the poorest countries in the world. 62 % of the employed Malian population lives from agriculture and livestock farming and generates 36 % of the domestic GDP (WorldBank, 2023a,b). Mali has different climatic conditions that also shape the agricultural production. The north is mostly characterised by desertification and is therefore less suitable for agricultural production. As a result, most agricultural production takes place in the south regions (FAO, 2015). The most common form is rain-fed agriculture for self-sufficiency. However, the dependence on agriculture makes Mali vulnerable to the effects of climate change. In 2020, the poverty headcount ratio at national poverty lines was 41.2 %. In other words, almost half of the Malian population lives below the poverty line (WorldBank, 2023c).

Our main result is that different types of client clusters in the agricultural sector in Mali exist. The loan repayment performance of these clusters differs as each group has different needs. Our work implies that different financial instruments are needed to strengthen the resilience of different groups of farmers in Mali. Accordingly, specific programs can be tailored for each of the clusters e.g., to ensure an effective and economically rewarding production, for food security and rural development programs of smallholders and family-based farmers. This research is therefore relevant for policymakers and the banking sector because policy and communication design need to consider these different types of customers comprehensively to be most efficient.

To the best of our knowledge, this is the first study analysing farmers repayment performance from a Malian commercial lender without a social mission in particular based on the foundation of an unsupervised clustering. Also, this research is the first to investigate credit repayment behaviour of farmers in combination with machine learning and regression techniques based on a unique and comprehensive data set derived from the Banque Nationale de Développement Agricole (BNDA). This estimation strategy combines theoretical considerations with data-driven elements, and therefore also contributes to the methodological literature.

The rest of the paper is structured as follows. Section 2 introduces the data, Section 3 presents the methodology. In Section 4 we presents our results and discuss them. Section 5 focuses on some concluding remarks.

2 Data

Unique and comprehensive loan data, which is used for our analysis is provided by a commercial Malian bank, the Banque Nationale de Développement Agricole (BNDA). The loan data covers all granted credits of the BNDA covering the period from January 2010 to April 2022. The BNDA started as an agricultural bank but has expanded its services over time. Today, the BNDA provides services for all sectors. The BNDA operates in seven out of ten regions (Gao, Kayes, Koulikoro, Mopti, Ségou, Sikasso, Tombouctou) and the capital district (Bamako) in Mali (BNDA, 2021). These are also the Malian regions where most of the agricultural production takes place (FAO, 2015).

Our study focuses only on borrowers who are classified as farmers. Agricultural loans are identified by the industry in which the credit customer is active. The data is extracted from the management information system (MIS) of the BNDA. The raw data contains 26,617 credits of 5,773 clients who are identified as farmers. Hence, this is a selected sample that is not representative of the general population, but rather comparable to other commercial financial institutes operating in Mali. As part of the credit scoring process, a number of borrower characteristics are collected, which include, for example, information on occupation, the age, the gender or the marital status. This customer information is entered manually into the system during the process of requesting the credit. In contrast, the credit information, such as the credit amount, the interest rate, the day of repayment, are generated automatically by the MIS. As part of our data cleaning, we had to exclude observations with missing values for certain variables (such as age, gender, or marital status) in order to conduct our clustering analysis. This was necessary because the bank employees responsible for manual data entry of the personal client information sometimes failed to provide all the required information, resulting in missing data that had to be excluded from our sample. Moreover, our analysis of repayment performance focuses only on loans that have already been repaid. Therefore, we have to exclude unfinished, open loans from our analysis and only focus on fully repaid loans.

For the clustering analysis, we aggregate all granted credit data at the client level. This lead to a sample of 3,335 customers. For the analysis of repayment performance only repaid credits are considered. The final sample for the analysis of the repayment performance consists of 3,335 customers and 9,469 credits granted and repayed credits of customers who work in the Malian agricultural sector and have received a credit from the BNDA between January 2010 and April 2022. The final sample is further described in section 4.1.

3 Estimation Strategy

In the first part (3.1) of this section we present the methodological steps of our cluster analysis and in the second part (3.2) we discuss how we measure loan repayment.

3.1 Clustering

Partitional clustering methods (such k-means and Partitioning Around Medoids (PAM)) are unsupervised machine learning techniques used to group a set of objects into meaningful and useful clusters such that the objects within the cluster are similar to each other and different from objects in other clusters. In other words, an advantage of these clustering techniques is that multiple variables help identify clusters with high similarity within groups that are dissimilar to other clusters (Chan & Mátyás, 2022; Lesmeister, 2015). The guiding principle of clustering analysis is to minimize intragroup variability and maximizing intergroup variability by some metric, e.g., distance connectivity, meanvariance, etc. (Ramasubramanian & Singh, 2017). For this study, cluster analysis is essentially about discovering farmer groups in Mali based on their credit behaviours. Through this, we can describe the structure of the data, and attempt to identify a logical organization of the data that simplifies the analysis. In other words, we adopt an exploratory approach to identify distinct clusters among Malian farmers.

Among the clustering methods, k-means and PAM are the most common methods. Both k-means and PAM method belongs to partitional clustering methods (Bishop & Nasrabadi, 2006). Partitional clustering decomposes a data set into a set of disjoint clusters. Given a data set of n points, a partitioning method constructs k ($k < n$) partitions of the data, with each partition representing a cluster. It classifies the data into k groups by satisfying the following requirements: a) each group contains at least one point, b) each point belongs to exactly one group (Lesmeister, 2015). Partitioning algorithms are more suitable for larger data sets as they are less computationally expensive. Therefore, partitioning algorithms are generally preferred for analysing large data sets (Bishop & Nasrabadi, 2006).

k-means clustering can, however, only analyse continuous quantitative variables. In contrast, PAM can process mixed data, both quantitative and qualitative, including nominal, ordinal, and interval ratio data (Lesmeister, 2015). In k-means clustering, each cluster is represented by the center or means of the data points belonging to the cluster. In PAM clustering, each cluster is represented by one of the objects in the cluster. As Graskemper et al. (2021) pointed out that the center of a cluster for k-means is not necessarily one of the input data points, but PAM chooses data points as centers and can be used distances arbitrarily. This is another advantage of PAM. Because a mean is easily influenced by extreme values, so that k-means clustering algorithm is sensitive to outliers. While PAM uses an actual point in the cluster to represent the center point, which is more robust to noises and outliers, so that it is more suitable when the data set contains outliers or noise (Kaufman & Rousseeuw, 1990).

The Gower coefficient, which measures the dissimilarity of all the observations to the nearest medoid, compares cases pairwise and calculates a dissimilarity between $farmer_i$ and $farmer_j$, which is essentially the weighted mean of the contributions of each variable (Lesmeister, 2015). It is defined as follows:

$$S_{ij} = \frac{\text{sum}\left(W_{ijk} \cdot S_{ijk}\right)}{\text{sum}\left(W_{ijk}\right)} \tag{1}$$

Here, S_{ijk} is the contribution provided by the kth variable and W_{ijk} is 1 if the kth variable is valid, or else 0. And,

$$\mathrm{S}_{ijk} = 1 - \left(\left|x_{jk} - x_{ik}\right|\right)/r_k \tag{2}$$

where r_k is the range of values for the kth variable.

After we defined the Gower coefficient (dissimilarity of all the observations to the nearest medoid), we use the Ward distance to minimize the dissimilarity (Lesmeister, 2015). Ward method minimizes the total within-cluster variance:

$$\sum_{i=1}^{k} \sum_{j=1}^{|C_i|} \text{dist}\left(x_j, \text{center}(i)\right) \tag{3}$$

where $|C_i|$ is the number of points in cluster i, $dist(x_j, center(i))$ is the Ward distance between point x_j and $center(i)$.

As the number of clusters is open, the selection of an optimal number of clusters is key to the results. There are many selection methods in the literature (Lesmeister, 2015; Graskemper et al., 2021). A smart initialization of the centroids in PAM is very important to obtain a good solution. A lesser number of clusters results in a rough sketch of the documents, while a larger number of clusters k results in comprehensive but mixed content. One method to find the appropriate number of clusters is to calculate the within-cluster variance with respect to the different k (Ghatak, 2017). This study mainly used the Elbow method to judge the optimal number of clusters with Eq. (3). Graphing the within-cluster variance by the clusters against the number of clusters, the point of decline of the marginal gain of added information reveals the optimal number of clusters (Ghatak, 2017). This number is independent from the researchers' opinion of the optimal number of clusters, as Graskemper et al. (2021) pointed out.

The analysis is conducted by using R statistics software.

3.2 Loan repayment

Different approaches to measuring loan repayment performance exist. Brehanu & Fufa (2008) apply a two-limit Tobit model to analyse the determinants of repayment rate of loans among smallholder farmers in Ethiopia. The authors use the proportion of loan repaid computed by dividing the amount of loan repaid to the total amount borrowed from the credit, ranging between between 0 and 1, as dependent variable. Nawai & Shariff (2012) use a mixed method approach and categorise the loan repayment of Malaysian microfinance clients in three categories (on time, late or defaulted). Weber & Musshoff (2012) use the proportion of credit with delinquent payments to analyse repayment behaviour for microcredit clients in Tanzania, while Pelka et al. (2015) use a linear probability models and a sequential logit model and measure loan repayment as the number of loan instalments in arrears for a microfinance institution in Madagascar.

The described approaches have in common that they attempt to capture the timeliness or delay of payments. The data set provided by the BNDA does not include delinquency payments. Based on our variable availability, we can only see whether a loan has been repaid on time or delayed, as well as the number of days the repayment was delayed. Therefore, similar to the measure of repayment performance employed by Nawai & Shariff (2012), we categorise the granted and repaid credits as on time or delayed.

The repayment behaviour can be adjusted by setting different time thresholds that define a delayed loan repayment. We follow the thresholds introduced by Weber & Musshoff (2017) and further differentiate a late repayment: DR1 indicates that the repayment is delayed by more than one day, DR15 if the repayment is missed by more than 15 days, and DR30 if the repayment is missed by more than 30 days. The DR1 is more sensitive in comparison to DR15 and DR30 since it captures delays starting from the first day. DR15 and DR30 capture long-term delays. We define a loan repayment as on time if the loan is repaid without delays or even before the last instalment (Sarwosri et al., 2016).

We start by using the continuous variable number of overdue days as regressand. We follow Weber & Musshoff (2012) and estimate an OLS regression of the following form:

$$Y_{i,y} = \beta_0 + \beta_1 c_i + \beta_2 l_{i,y} + \delta_b + \delta_t + \epsilon_{i,y} \tag{4}$$

where $Y_{i,y}$ is the number of overdue days for individual i and credit y, c_i is a vector containing client information and $l_{i,y}$ contains loan information for individual i and credit y. $\epsilon_{i,y}$ is the independently and identically distributed error term with a mean of zero and a variance of σ_e^2. We use branch specific δ_b and time δ_t fixed effects.

Additionally, we follow the estimation strategy of De Andrés et al. (2021) and Grohmann et al. (2020) and estimate a logit regression with heteroscedasticity robust standard errors and employ the different indicators of delayed repayments as regressand. A logistic regression model is typically used to model the relationship between a binary dependent variable and one or more independent variables (Hansen, 2022). We estimate the following form:

$$P(Default_{i,y} = 1|x) = \beta_0 + \beta_1 c_i + \beta_2 l_{i,y} + \delta_b + \delta_t + \epsilon_{i,y}) \tag{5}$$

where the dependent variable is a dummy equal to 1 if a credit was repaid in time as well as a dummy indicating the severity of delayed repayments (DR1, DR15, and DR30), and 0 otherwise for individual i and credit y. c_i is a vector containing client information and $l_{i,y}$ contains loan information for individual i and credit y. $\epsilon_{i,y}$ is the the error term for unobserved heterogeneity, affecting the outcome and is assumed to be uncorrelated with the independent variables in these models. Again we use branch specific δ_b and time δ_t fixed effects. The analysis is conducted by using Stata 17.

In short, we combine the suggested estimation strategies by Weber & Musshoff (2012), De Andrés et al. (2021), and Grohmann et al. (2020) to examine the loan repayment for each cluster within our sample of agricultural loans disbursed by the commercial bank BNDA.

4 Results and discussion

In this section, we provide a detailed description of the finale data set in subsection 4.1. Additionally, we define the clusters in subsection 4.2 based on PAM and the Elbow method and provide a comprehensive overview of each cluster based on the summary statistics for each cluster. We then estimate the loan repayment of each cluster in subsection 4.3 to investigate whether there are any differences in repayment performance between the clusters, and identify the regressors that explain the different repayment behaviours.

4.1 Sample description

Table 1 shows the summary statistic of the variables that are included in our clustering analysis. A number of borrower characteristics such as age, gender and marital status are added. We select the three socio-demographic variables based on a comprehensive desk research, stating that the farmers' age (De Lauwere, 2005; Fafchamps et al., 2014), gender (Lambrecht et al., 2018; Khandker & Koolwal, 2016; Beaman & Dillon, 2018; De Mel et al., 2009; Chamboko et al., 2021), and marital status (Graskemper et al., 2021; Weber & Musshoff, 2017) can have a considerable effect on farmers agricultural behaviour as well as on the financial behaviour in general. Furthermore, several variables related to the individual credit behaviour (i.e. mean credit duration of all granted credits, mean interest rates of all granted credits, the frequency of the credit, the number of granted credits by customer as well as the share of each month when the credit was granted) are included in the table as well as in the clustering process.

The average age of the farmers is 41 years and the majority is married (58 %) and male (91 %). The interest rate varies between 0 and 15 %. The BNDA offers certain specific financial products with an interest rate of 0 % (BNDA, 2021). The average granted loan size is about 2,031,681 CFA-Franc. The number of granted credits per farmer varies between 1 and 27. The mean number of credits per farmer is 2.84, while the mean credit duration is around 34 months. A majority of the granted credits have a monthly repayment (59 %) and were granted in the southern regions of Mali (98 %). This are also the regions were most of the agricultural production takes place (FAO, 2015).

4.2 Definition of clusters

Based on PAM and the Elbow method, three groups of clusters are identified. Figure 1 plots the sum of within-cluster variance with respect to the number of clusters. The figure shows that a higher value of k will reduce the variance within clusters. The point at which the marginal gain in additional information decreases shows the optimal number of clusters (Ghatak, 2017). The pronounced "bend" observed in Figure 1, indicated that adding another cluster does not substantially reduce variance. This "bend" is the so-called "elbow". Focusing on the steepest turnover in Figure 1, one can see that the optimal number of clusters is three.

In other words, the ML technique created 3 groups and allocated all farmers in our sample (Observations = 3,335) in these three groups. Based on figure 2, the clusters can be defined as follows:

(1) ”Frequent lowest-cost farmers": Represent 23.3% of the sample. The mean age lies between the other two clusters. The cluster is characterized by the largest number of granted credits, a relatively large share of female farmers, the lowest credit value disbursed, the lowest interest rate and lowest credit duration. In short, it represents farmers who require relatively frequent

Table 1: Summary Statistic

	Full Sample			
	Mean	SD	Min.	Max.
Age of customer (continuous variable)	41.06	15.02	14.00	94.00
Gender of customer	0.91	0.28	0.00	1.00
Dummy if customer is married (0= no, 1= yes)	0.58	0.49	0.00	1.00
Dummy if credit granted in southern regions (0= no, 1= yes)	0.98	0.13	0.00	1.00
Number of granted credits by client (continuous variable)	2.84	3.14	1.00	27.00
Credit amount disbursed (continuous variable)	2.03e+06	4.17e+06	25,000.00	1.00e+08
Mean credit amount disbursed (continuous variable)	2.02e+06	3.75e+06	40,000.00	9.99e+07
Max. credit amount disbursed (continuous variable)	2.56e+06	5.40e+06	40,000.00	1.00e+08
Min. credit amount disbursed (continuous variable)	1.64e+06	2.94e+06	25,000.00	9.99e+07
Credit duration (in month)	34.28	52.44	1.00	1,418.00
Mean credit duration (in month)	27.83	18.42	1.00	120.00
Min. credit duration (in month)	24.67	20.03	1.00	120.00
Max. credit duration (in month)	32.71	18.58	1.00	120.00
Interest rate of the granted credit	9.06	3.12	0.00	15.00
Mean interest rate of the granted credit	9.10	2.62	1.00	14.50
Min. interest rate of the granted credit	7.91	3.69	0.00	14.00
Max. interest rate of the granted credit	9.96	2.75	1.00	15.00
Dummy if frequency of credit is monthly (0= no, 1= yes)	0.59	0.49	0.00	1.00
Share of credits granted in January	0.04	0.17	0.00	1.00
Share of credits granted in February	0.05	0.19	0.00	1.00
Share of credits granted in March	0.11	0.26	0.00	1.00
Share of credits granted in April	0.16	0.32	0.00	1.00
Share of credits granted in May	0.20	0.34	0.00	1.00
Share of credits granted in June	0.11	0.26	0.00	1.00
Share of credits granted in July	0.09	0.23	0.00	1.00
Share of credits granted in August	0.06	0.18	0.00	1.00
Share of credits granted in September	0.05	0.15	0.00	1.00
Share of credits granted in October	0.05	0.16	0.00	1.00
Share of credits granted in November	0.04	0.15	0.00	1.00
Share of credits granted in December	0.04	0.15	0.00	1.00
Share of credits granted related to agriculture	0.08	0.25	0.00	1.00
Share of credits granted related to employment	0.44	0.47	0.00	1.00
Share of credits granted related to other reasons	0.38	0.47	0.00	1.00
Observations	3,335			

Notes: To increase the transparency, the summary statistics, display the continuous and logarithmic values. The continuous variables credit amount disbursed and the interest rate are used in the logarithmic form. The number of due dates and the number of granted credits by client remain in the continuous form.

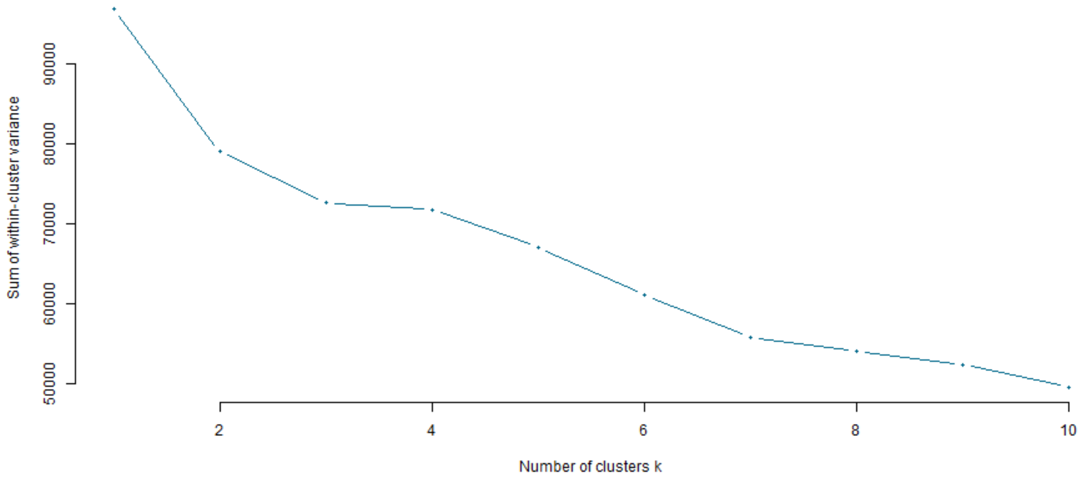

Figure 1: *Number of clusters*

smaller loans for less capital-intensive activities. We will define this cluster as ”Frequent lowest-cost farmers” (FL-CF).

(2) ”Experienced farmers”: Represent 37.3% of the sample. The cluster is characterized by the largest mean age and the largest number of female farmers, the number of granted credits lies between cluster 1 and cluster 3, the credit value disbursed and the mean credit duration range between cluster 1 and cluster 3. This cluster has the largest mean interest rate. We will define this cluster as ”Experienced farmers” (EF).

(3) ”High-volume, long-term farmers”: Represent 39.4% of the sample. The cluster has the lowest mean age and the lowest number of female farmers. The cluster is characterized by the lowest number of granted credits, the largest credit value disbursed, the largest credit duration, which suggests that it represents farmers who require larger loans for longer-term investments. The interest rates of this cluster range between cluster 1 and cluster 2 (EF), which could indicate that this group has a more diverse range of credit needs. We will define this cluster as ”High-volume, long-term farmers” (HV-LTF).

The three different clusters that we have identified with PAM and the Elbow method differ in various farmers characteristics (e.g., age or share of female farmers) or the characteristics of received credits (e.g., interest rate or granted loan amount) (compare Figure 2). Hence, we can conclude that different clusters of agricultural credit clients exist in Mali.

Based on the summary statistic for each cluster (compare Appendix Table A1), we further describe the characteristics in detail. HV-LTF are the youngest farmers within the sample. EF are the oldest but close to FL-CF. However the minimum and maximum values for each of the three clusters has a relatively similar range (ranging from 14 to 16 and from 85 to 94 years). Also, in terms of marriage, FL-CF and EF have the highest share of married farmers, while HV-LTF has the lowest share of married farmers. FL-CF and EF have a similar ratio of male and female farmers,

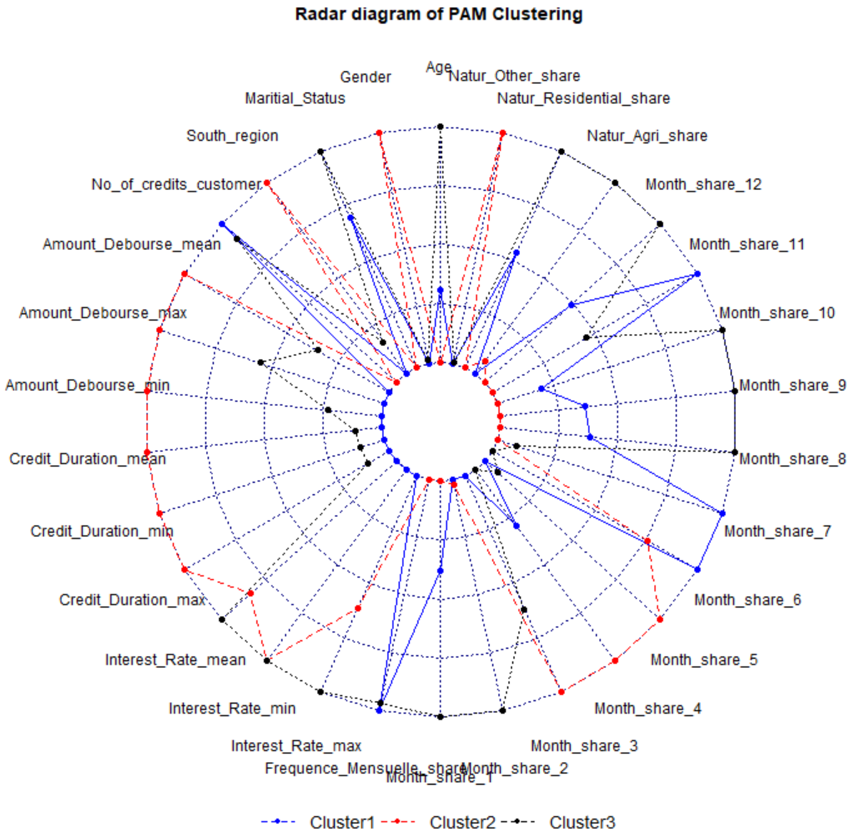

Figure 2: *PAM results*

Notes: Characteristics of different farmer credit clients clusters. The figure show the relative distribution of the expression of the variables. The inner circle indicates a low expression and the outer circle a high expression.

while HV-LTF includes almost completely male farmers. 15.5 % of farmer in cluster FL-CF are female, for cluster EF this share is 16.5 % and for HV-LTF it is 0 %. Having in mind that 8.5 % of the complete farmers are female, we can conclude that these female farmers are allocated in two clusters.

FL-CF has with on average more than 5 the largest number of granted credits, while EF has a mean of 2.5 and HV-LTF of less than 2 granted credits. In contrast, FL-CF has the lowest mean value of granted and disbursed credit, while HV-LTF has the highest mean value of disbursed credit. EF ranges between FL-CF and HV-LTF. HV-LTF has with about 3.5 years the highest mean duration of granted credits, while FL-CF has with about 1 year the lowest mean credit duration and EF is with about 1.4 years closer to FL-CF. EF has with more than 10 % the largest mean interest rate, while FL-CF has the lowest (less than 6 %) and HV-LTF has an interest with more than 9 close to the mean interest rate of EF. Almost all credits of FL-CF and EF are repaid on a monthly basis, while the majority of the granted credits of HV-LTF are not repaid monthly and hence repaid annually, bi-annually, or irregularly. Focusing on the timing of the credits, we can see that on average the highest number of credits were granted in July for FL-CF, EF received credits all around the year with no clear trend, while HV-LTF received the majority of the credits in May. EF has the highest share of credits granted for activities directly associated with agricultural inputs (e.g., buying seed, fattening). EF has the largest share of short- and medium-term credits related to employment issues. HV-LTF has the largest share of credits that are related to other reasons that include equipment.

4.3 Loan repayment of the clusters

In the following we estimate the repayment performance for each cluster. Table A2 in the Appendix shows the descriptive statistics for the analysis of the credit repayment performance. Cluster 1 (Fl-CF) has the lowest mean overdue days (8.3), followed by cluster EF (13.8) and cluster HV-LTF (36.4). The cluster EF has the highest share of credits repaid in time (39%), followed by cluster FL-CF (35%) and cluster HV-LTF (18%). Focusing on the indicators that highlight the magnitude of the delayed payment, one can see that cluster Fl-CF has the largest mean of credits delayed by more than one day (DR1) (55%), followed by cluster EF (50%) and cluster HV-LTF (43%). The indicators DR15 and DR30 capture long-term delays. Also for DR15 cluster FL-CF has the largest mean (26%), followed by cluster EF (23%) and cluster HV-LTF (21%). However, when focusing on DR30 the indicator the most sensitive to long-tern delayed credit repayments, cluster HV-LTF has the largest mean (15%), while cluster FL-CF (0%) and EF (3%) have only a very low mean for this indicator. This indicates that if cluster FL-CF or cluster EF have a delayed credit it is more likely that these a credit are delayed only in the short- and medium-run. While for cluster 3, it is more likely that such a credit is delayed in the medium- or long-run.

As shown in the previous subsection, the three clusters differ in their personal characteristics, credit characteristics and credit demand behaviour. Table 2 highlight the repayment performance of each cluster, based on the OLS and logit regressions. In column (1) to (3) the dependent variable is a dummy indicating if a credit was prepaid in time. In column (4) to (6) the number of overdue days is the dependent variable.

Focusing on column (1) to (3), we can see that for cluster FL-CF and EF, a robust statistically significant negative association between the credit amount disbursed and the dummy if credit was repaid in time exist. We observe a positive but statistically insignificant for cluster HV-LTF. Only for cluster EF and HV-LTF, a positive relationship between the interest rate and the dummy if credit was repaid in time exist, while the effect for cluster FL-CF is negative statistically significant associated. For cluster FL-CF and EF a negative statistically significant relationship between the number of granted credits and the dummy if credit was repaid in time exist. However, this relationship is positive and statistically significant for cluster HV-LTF. Being a male farmer is statistically significant positive associated with a repayment in time for cluster EF and HV-LTF, and statistical significant negative for cluster FL-CF.

Focusing on columns (4) to (6), where the number of overdue days is the dependent variable, we can see that for cluster FL-CF, EF and HV-LTF a positive and statistically significant relationship between the loan amount disbursed and the number of days the repayment was delayed exist. Focusing on the effect of interest rates, we observe heterogeneous effects. While we see a statistically significant positive association for cluster FL-CF, we see a statistically significant negative association between the interest rates for cluster HV-LTF and statistically insignificant association for cluster EF and the number of days the repayment was delayed. Focusing on cluster EF and HV-LTF, the number of granted credits is statistically significant negatively associated with the number of days the repayment was delayed and statistically significant positively associated for cluster FL-CF. The number of due dates is for all three clusters statistically significant negatively associated with the number of days the repayment was delayed. Being a male farmer is statistically significant positively associated with the number of days the repayment of the credit is delayed for cluster FL-CF. This association is statistical insignificant for cluster EF and HV-LTF.

Table 3 focuses on the analysis of the delayed repayment behaviour based on the different thresholds used by Weber & Musshoff (2017) and further differentiate a late repayment. The dependent variable in columns (1) to (3) is a binary variable represented by a dummy variable, which indicates whether the repayment was delayed by at least one day (DR1). For columns (4) to (6), the dependent variable is also represented by a dummy variable, but it indicates whether the repayment was delayed by at least 15 days (DR15). In column (7) to (9), the dependent variable is a binary variable represented by a dummy variable that indicates whether the repayment was missed by at least 30 days (DR30).

Focusing on columns (1) to (3), we can observe that for all cluster a robust statistically significant positive association between the credit amount disbursed and the dummy if a credit was repaid with a delay of more than one day exist. For cluster 2 (EF) and 3 (HV-LTF) we observe a statistically significant negative association between the interest rate and the delayed repayment. This association is significant negative positive for cluster FL-CF. For cluster 1 (FL-CF) and 2 (EF) a positive statistically significant relationship between the number of granted credits and the dummy indicating the delayed repayment exist. However, this relationship is negative and statistically significant for cluster 3 (HV-LTF).

Regarding columns (4) to (6) where the dependent variable is a dummy if the repayment is delayed by at least 15 days, we can see for all three cluster a robust statistically significant positive association between the credit amount disbursed and the dummy if a credit was repaid with a delay of more than 15 days exist. For cluster FL-CF and EF, we observe a statistically significant

Table 2: Logit results

Dependent variable:	Dummy if credit repaid in time			Number of overdue days		
	(1) FL-CF	(2) EF	(3) HV-LTF	(4) FL-CF	(5) EF	(6) HV-LTF
Credit amount disbursed (logarithmic value)	-0.850***	-0.361***	0.138	3.122***	3.988***	20.301***
	(0.000)	(0.000)	(0.382)	(0.000)	(0.000)	(0.000)
Interest rate (logarithmic value)	-0.267*	2.240***	4.847	0.362	3.752	64.899***
	(0.081)	(0.001)	(0.349)	(0.493)	(0.858)	(0.001)
Number of due dates (in month, logarithmic value)	0.518***	0.144**	-0.545	-2.608***	-11.334***	-24.763***
	(0.000)	(0.037)	(0.200)	(0.000)	(0.000)	(0.002)
Credit duration (in month)	0.010***	0.005***	0.001	-0.035***	0.022	0.007
	(0.000)	(0.001)	(0.183)	(0.001)	(0.744)	(0.941)
Gender of customer	-0.216*	0.257**	0.925	0.847*	0.731	22.307
	(0.061)	(0.036)	(0.347)	(0.062)	(0.817)	(0.501)
Number of granted credits by client (continuous variable)	-0.074***	-0.033**	0.102**	0.225***	-2.376***	-6.938***
	(0.000)	(0.023)	(0.019)	(0.000)	(0.000)	(0.000)
Branch FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.154	0.073	0.353	0.103	0.060	0.236
Observations	4,124	3,145	2,093	4,132	3,146	2,191

Notes: In column (1) to (3) where we estimate a logit regression the dependent variable is a dummy indicating if a credit was repaid in time. The dependent variable in column (4) to (6) where we estimate an OLS regression is the number of overdue days (continuous variable). As can be seen from the number of observations, the sample size for the logit regression (column (1) to (3)) is slightly reduced because Stata excludes observations that could lead to a potential problem, so that the remaining coefficients in the model are not biased. Stata then fit the model using the remaining observations. p-values in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

negative relationship between the number of due dates and the delayed repayment. However, this relationship is statistically insignificant for cluster (FL-CF). Heterogeneous associations between the number of granted credits and the dummy indicating the delayed repayment exist, while we can not see any statistically significant associations between gender and DR 15.

As for columns (7) to (9) where the dependent variable is a dummy if the repayment is delayed by at least 30 days, we can see that for cluster EF (EF) and HV-LTF a robust statistically significant positive association between the credit amount disbursed and the dummy if a credit was repaid with a delay of more than 30 days exist. This association is statistically insignificant for cluster FL-CF (FL-CF). For cluster FL-CF and EF we observe a statistically significant positive association between interest rate and DR30. The number of due dates are statistically significant negative associated with DR30 for cluster EF and HV-LTF, and statistically insignificant for cluster FL-CF. For cluster EF and HV-LTF a negative statistically significant relationship between the number of granted credits and the dummy indicating the delayed repayment exist. However, this relationship is negative and insignificant for cluster FL-CF.

In conclusion, a higher loan volume is on average associated with a higher risk of a delayed loan repayment. This result is robust for all three clusters across most of the different delayed repayment indicators. We do not find any robust association between gender and a delayed repayment. Heterogeneous effects of interest rates and the number of granted credits are observed. The number of due dates are across all clusters negatively associated with delayed repayment indicators. Our findings emphasize the need for a heterogeneous perspective of different groups of farmers and of their credit behaviour in general.

Placing our results in a larger context, the literature shows that MFIs with more women in their lending portfolios have lower default risk and higher repayment rates (D'espallier et al., 2011). Our results show a nuanced picture. Female farmers are mostly clustered in two out of three clusters (FL-CF and EF). In cluster FL-CF being a male is on average negatively associated with a loan repayment in time, while the effect is the opposite for cluster EF. When looking at the disbursed loan amount we find significantly higher associations between loan amount and delayed repayment (with decreasing effects for cluster FL-CF and increasing effects for cluster HV-LTF). This results are in line with Weber & Musshoff (2012) who find a higher credit risk for larger loans. The interest rate is also associated with a delayed repayment performance across different specifications and clusters. We find a positive association for some clusters and a negative association for others. This differentiated result is consistent with previous studies such as (Raghunathan et al., 2011) who find that higher interest rates are associated with better repayment performance for MFi borrower groups, and those that find the opposite (Banerjee, 2013). However, when drawing conclusions from the results of this study, it is important to remember that it is an observational study.

5 Conclusion

The formal credit sector continues to play an important role in financing the transformation towards resilient and sustainable development. Hence, an unsupervised grouping of these loan clients in the agricultural sector is of interest in improving policy recommendations, adjust market failures and hence, improving the provision of liquidity to the agricultural sector in Mali. The objective of this paper was (1) to provide a depiction of commercial bank clients in the agricultural sector in

Table 3: Logit results

Dependent variable:	DR1			DR15			DR30		
	(1) FL-CF	(2) EF	(3) HV-LTF	(4) FL-CF	(5) EF	(6) HV-LTF	(7) FL-CF	(8) EF	(9) HV-LTF
Credit amount disbursed (logarithmic value)	0.856***	0.359***	0.569***	0.592***	0.287***	0.976***	0.276	0.486***	1.112***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.347)	(0.000)	(0.000)
Interest rate (logarithmic value)	0.077	-5.477***	-2.719	-0.061	-3.689***	-3.953	2.588***	2.125	2.801***
	(0.546)	(0.000)	(0.552)	(0.584)	(0.000)	(0.328)	(0.000)	(0.242)	(0.000)
Number of due dates (in month, logarithmic value)	-0.766***	-0.451***	0.586*	-0.477***	-0.452***	-0.264	-0.272	-1.190***	-0.977***
	(0.000)	(0.000)	(0.068)	(0.000)	(0.000)	(0.473)	(0.635)	(0.000)	(0.001)
Credit duration (in month)	-0.009***	-0.008***	-0.001	-0.006***	-0.007**	-0.001	-0.007	-0.002	-0.002
	(0.002)	(0.000)	(0.555)	(0.002)	(0.016)	(0.718)	(0.244)	(0.485)	(0.442)
Gender of customer	0.203*	-0.102	-0.743	0.191	-0.068	0.283	0.000	0.989	0.617
	(0.061)	(0.402)	(0.541)	(0.103)	(0.632)	(0.852)	(.)	(0.122)	(0.726)
Number of granted credits by client (continuous variable)	0.078***	0.032**	-0.167***	0.044***	0.001	-0.453***	-0.015	-0.315***	-0.469***
	(0.000)	(0.024)	(0.001)	(0.000)	(0.938)	(0.000)	(0.886)	(0.000)	(0.000)
Branch FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.120	0.082	0.455	0.069	0.057	0.378	0.154	0.259	0.342
Observations	4,127	3,145	2,154	4,124	3,132	2,148	1,297	2,602	2,167

Notes: Dependent variable in column (1) to (3) a dummy indicates if repayment is delayed by more than one day, in column (4) to (6) a dummy indicates if repayment is delayed by more than 15 days, and in in column (7) to (9) a dummy indicates if repayment is delayed by more than 30 days. As can be seen from the number of observations, the sample size for the logit regressions is reduced because Stata excludes observations that could lead to a potential problem, so that the remaining coefficients in the model are not biased. Stata then fit the model using the remaining observations. p-values in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Mali and (2) to estimate whether the loan repayment performance of the clusters differ. Thereby, we contribute to the understanding of credit clients and their needs. Our analysis show the heterogeneity of farmers and their credit behaviour in Mali. The unsupervised ML based on a large pool of variables created a comprehensive overview of different clusters within the agricultural credit market in Mali. The analysis of the repayment performance highlighted that for each cluster different variables are associated with the repayment performance. In dependency of the cluster, some regressors even have a reverse relationship (e.g. number of granted credits on DR1, DR15, and DR30). This highlights an important advantage of our analysis and we highlighted associations that one might not have seen without the clustering of the customers.

Our suggestions for policymakers are that different instruments are needed to strengthen the resilience of different groups of farmers in Mali. Our analysis showed that repayment performance varies across clusters and for each indicator of delayed repayments (DR1, DR15, and DR 30) different associations exist. For example, the most severe long-term repayment delays occur almost exclusively in one cluster (HV-LTF). Consequently, specific customer clusters that can be identified based on unsupervised ML are those with a high probability of long-term delayed repayments. The main variables affecting DR30 are loan amount, interest rates and the number of granted credits per client. Financial institutions seeking to improve their clients' repayment performance should focus most on these three variables. However, if an MFI is competitive with its interest rates, then it will be more difficult to change the interest rate, as (Raghunathan et al., 2011) has already noted. Our findings are an important prerequisite for the design of future agricultural policies and financial products for the agricultural sector to increase the overall resilience of the sector and to contribute to the reduction of poverty in Mali, in general. This research is therefore relevant for policymakers to enable more evidence-based agricultural and food policies as well as in the banking sector (El Benni et al., 2023).

Our paper contributes to a broader debate: the analysis of credit repayment in the Global South, and secondly, the use of ML techniques in agricultural economics. The analysis highlights that ML is a useful extension to the analysis of credit data and the business decision-making process of a company. This paper contributes to previous studies that highlighted how ML is applied in business decision-making processes. Our research focuses only on the traditional finance sector. Therefore, further research on non-traditional finance using causal methods is encouraged. On the one hand, a strength of the study is the use of data from multiple years and production seasons. On the other hand, the data only includes granted loans and no information regarding credit rationing. Including rejected loan applications would also be of interest to build clusters of farmers excluded from the traditional financial sector. As our results are based on a commercial bank in Mali, our approach could also be applied to other countries in the Global South to contribute to the heterogeneous understanding of the Global South.

References

Banerjee, A. V. (2013). Microcredit under the microscope: What have we learned in the past two decades, and what do we need to know? *Annu. Rev. Econ.*, *5*, 487–519.

Banerjee, A. V., & Duflo, E. (2014). Do firms want to borrow more? testing credit constraints using a directed lending program. *Review of Economic Studies*, *81*, 572–607.

Barry, P. J. (2001). Modern capital management by financial institutions: implications for agricultural lenders. *Agricultural Finance Review*, *61*, 103–122.

Bartkowski, B., Schüßler, C., & Müller, B. (2022). Typologies of european farmers: approaches, methods and research gaps. *Regional Environmental Change*, *22*, 1–13.

Batung, E. S., Mohammed, K., Kansanga, M. M., Nyantakyi-Frimpong, H., & Luginaah, I. (2023). Credit access and perceived climate change resilience of smallholder farmers in semi-arid northern ghana. *Environment, Development and Sustainability*, *25*, 321–350.

Beaman, L., & Dillon, A. (2018). Diffusion of agricultural information within social networks: Evidence on gender inequalities from mali. *Journal of Development Economics*, *133*, 147–161.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* volume 4. Springer.

BNDA (2021). Rapport annuel 2020.

Brehanu, A., & Fufa, B. (2008). Repayment rate of loans from semi-formal financial institutions among small-scale farmers in ethiopia: Two-limit tobit analysis. *The Journal of Socio-Economics*, *37*, 2221–2230.

Carter, C., Cui, X., Ghanem, D., & Mérel, P. (2018). Identifying the economic impacts of climate change on agriculture. *Annual Review of Resource Economics*, *10*, 361–380.

Chamboko, R., Cull, R., Gine, X., Heitmann, S., Reitzug, F., & Van Der Westhuizen, M. (2021). The role of gender in agent banking: Evidence from the democratic republic of congo. *World Development*, *146*, 105551.

Chan, F., & Mátyás, L. (2022). *Econometrics with machine learning*. Springer.

Channa, H., Ricker-Gilbert, J., Feleke, S., & Abdoulaye, T. (2022). Overcoming smallholder farmers' post-harvest constraints through harvest loans and storage technology: Insights from a randomized controlled trial in tanzania. *Journal of Development Economics*, *157*, 102851.

Cui, X., & Xie, W. (2022). Adapting agriculture to climate change through growing season adjustments: Evidence from corn in china. *American Journal of Agricultural Economics*, *104*, 249–272.

Daloğlu, I., Nassauer, J. I., Riolo, R. L., & Scavia, D. (2014). Development of a farmer typology of agricultural conservation behavior in the american corn belt. *Agricultural Systems*, *129*, 93–102.

De Andrés, P., Gimeno, R., & de Cabo, R. M. (2021). The gender gap in bank credit access. *Journal of Corporate Finance*, *71*, 101782.

De Lauwere, C. (2005). The role of agricultural entrepreneurship in dutch agriculture of today. *Agricultural Economics*, *33*, 229–238.

De Mel, S., McKenzie, D., & Woodruff, C. (2009). Are women more credit constrained? experimental evidence on gender and microenterprise returns. *American Economic Journal: Applied Economics*, *1*, 1–32.

Dumortier, J., Carriquiry, M., & Elobeid, A. (2021). Impact of climate change on global agricultural markets under different shared socioeconomic pathways. *Agricultural Economics*, *52*, 963–984.

D'espallier, B., Guérin, I., & Mersland, R. (2011). Women and repayment in microfinance: A global analysis. *World development*, *39*, 758–772.

El Benni, N., Grovermann, C., & Finger, R. (2023). Towards more evidence-based agricultural and food policies. *Q Open*, (p. qoad003).

Fafchamps, M., McKenzie, D., Quinn, S., & Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in ghana. *Journal of development Economics*, *106*, 211–226.

FAO (2015). Profil de pays – mali.

Faye, B., Webber, H., Gaiser, T., Müller, C., Zhang, Y., Stella, T., Latka, C., Reckling, M., Heckelei, T., Helming, K. et al. (2023). Climate change impacts on european arable crop yields: Sensitivity to assumptions about rotations and residue management. *European Journal of Agronomy*, *142*, 126670.

Flory, J. A. (2018). Formal finance and informal safety nets of the poor: Evidence from a savings field experiment. *Journal of Development Economics*, *135*, 517–533.

Fujimori, S., Wu, W., Doelman, J., Frank, S., Hristov, J., Kyle, P., Sands, R., Van Zeist, W.-J., Havlik, P., Domínguez, I. P. et al. (2022). Land-based climate change mitigation measures can affect agricultural markets and food security. *Nature Food*, *3*, 110–121.

Ghatak, A. (2017). *Machine learning with R*. Springer.

Ghosh, R., Sen, K. K., & Riva, F. (2020). Behavioral determinants of nonperforming loans in bangladesh. *Asian Journal of Accounting Research*, *5*, 327–340.

Gilbert, C. L., Christiaensen, L., & Kaminski, J. (2017). Food price seasonality in africa: Measurement and extent. *Food policy*, *67*, 119–132.

Godquin, M. (2004). Microfinance repayment performance in bangladesh: How to improve the allocation of loans by mfis. *World development*, *32*, 1909–1926.

Graskemper, V., Yu, X., & Feil, J.-H. (2021). Farmer typology and implications for policy design–an unsupervised machine learning approach. *Land Use Policy*, *103*, 105328.

Grohmann, A., Herbold, S., & Lenel, F. (2020). Repayment under flexible loan contracts: Evidence from tanzania. *DIW Berlin Discussion Paper*, *1884*, 1–40.

Guarín, A., Rivera, M., Pinto-Correia, T., Guiomar, N., Šūmane, S., & Moreno-Pérez, O. M. (2020). A new typology of small farms in europe. *Global Food Security*, *26*, 100389.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Hellmann, T. F., Murdock, K. C., & Stiglitz, J. E. (2000). Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough? *American economic review*, *91*, 147–165.

Hertel, T. W., & Rosch, S. D. (2010). Climate change, agriculture, and poverty. *Applied economic perspectives and policy*, *32*, 355–385.

IPCC (2022). Climate change 2022: Impacts, adaptation and vulnerability. contribution of working group ii to the sixth assessment report of the intergovernmental panel on climate change.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, *344*, 68–125.

Khandker, S. R., & Koolwal, G. B. (2016). How has microcredit supported agriculture? evidence using panel data from bangladesh. *Agricultural Economics*, *47*, 157–168.

Lambrecht, I., Schuster, M., Asare Samwini, S., & Pelleriaux, L. (2018). Changing gender roles in agriculture? evidence from 20 years of data in ghana. *Agricultural Economics*, *49*, 691–710.

Lesmeister, C. (2015). *Mastering machine learning with R*. Packt Publishing Ltd.

Methorst, R. R., Roep, D. D., Verhees, F. F., & Verstegen, J. J. (2017). Differences in farmers' perception of opportunities for farm development. *NJAS-Wageningen Journal of Life Sciences*, *81*, 9–18.

Nawai, N., & Shariff, M. N. M. (2012). Factors affecting repayment performance in microfinance programs in malaysia. *Procedia-Social and Behavioral Sciences*, *62*, 806–811.

Pelka, N., Musshoff, O., & Weber, R. (2015). Does weather matter? how rainfall affects credit risk in agricultural microfinance. *Agricultural Finance Review*, *75*, 194–212.

Raghunathan, U. K., Escalante, C. L., Dorfman, J. H., Ames, G. C., & Houston, J. E. (2011). The effect of agriculture on repayment efficiency: a look at mfi borrowing groups. *Agricultural Economics*, *42*, 465–474.

Ramasubramanian, K., & Singh, A. (2017). *Machine learning using R*. 1. Springer.

Rosenzweig, M. R. (2001). Savings behaviour in low-income countries. *Oxford review of economic policy*, *17*, 40–54.

Sarwosri, A. W., Römer, U., & Musshoff, O. (2016). Are african female farmers disadvantaged on the microfinance lending market? *Agricultural Finance Review*, *76*, 477–493.

Stephens, E. C., & Barrett, C. B. (2011). Incomplete credit markets and commodity marketing behaviour. *Journal of agricultural economics*, *62*, 1–24.

Suri, T., & Udry, C. (2022). Agricultural technology in africa. *Journal of Economic Perspectives*, *36*, 33–56.

Webber, H., Ewert, F., Olesen, J. E., Müller, C., Fronzek, S., Ruane, A. C., Bourgault, M., Martre, P., Ababaei, B., Bindi, M. et al. (2018). Diverging importance of drought stress for maize and winter wheat in europe. *Nature communications*, *9*, 1–10.

Weber, R., & Musshoff, O. (2012). Is agricultural microcredit really more risky? evidence from tanzania. *Agricultural Finance Review*, *72*, 416–463.

Weber, R., & Musshoff, O. (2017). Can flexible agricultural microfinance loans limit the repayment risk of low diversified farmers? *Agricultural Economics*, *48*, 537–548.

WorldBank (2023a). Agriculture, forestry, and fishing, value added (% of gdp) - mali. `https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=ML`. Accessed: 2023-02-22.

WorldBank (2023b). Employment in agriculture (% of total employment) (modeled ilo estimate) - mali. `https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS?locations=ML`. Accessed: 2023-02-22.

WorldBank (2023c). Poverty headcount ratio at national poverty lines (% of population) - mali. `https://data.worldbank.org/indicator/SI.POV.NAHC?locations=ML`. Accessed: 2023-02-22.

6 Appendix

Table A1: Summary statistics for farmer characteristics used in PAM

	Cluster 1 - FL-CF				Cluster 2 - EF				Cluster 3 - HV-LTF			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Age of customer (continuous variable)	41.90	13.28	14.00	85.00	43.78	13.35	16.00	89.00	37.99	16.84	16.00	94.00
Gender of customer	0.87	0.34	0.00	1.00	0.86	0.35	0.00	1.00	1.00	0.06	0.00	1.00
Dummy if customer is married (0= no, 1= yes)	0.67	0.47	0.00	1.00	0.68	0.47	0.00	1.00	0.43	0.50	0.00	1.00
Dummy if credit granted in southern regions (0= no, 1= yes)	0.97	0.17	0.00	1.00	0.97	0.17	0.00	1.00	1.00	0.04	0.00	1.00
Number of granted credits by client (continuous variable)	5.32	4.82	1.00	27.00	2.53	2.22	1.00	19.00	1.67	1.26	1.00	11.00
Credit amount disbursed (continuous variable)	625004.88	1.07e+06	30,000.00	1.50e+07	1.85e+06	5.41e+06	25,000.00	1.00e+08	3.04e+06	3.67e+06	150000.00	9.99e+07
Mean credit amount disbursed (continuous variable)	607522.60	771557.39	50,000.00	1.30e+07	1.76e+06	4.14e+06	40,000.00	6.46e+07	3.10e+06	4.08e+06	150000.00	9.99e+07
Max. credit amount disbursed (continuous variable)	1.39e+06	1.80e+06	50,000.00	2.65e+07	2.46e+06	6.87e+06	40,000.00	1.00e+08	3.36e+06	5.08e+06	150000.00	1.00e+08
Min. credit amount disbursed (continuous variable)	246008.74	421128.74	25,000.00	6.50e+06	1.22e+06	2.46e+06	25,000.00	3.00e+07	2.86e+06	3.66e+06	100000.00	9.99e+07
Credit duration (in month)	13.34	17.52	1.00	300.00	20.39	31.60	1.00	589.00	59.80	69.01	9.00	1,418.00
Mean credit duration (in month)	12.46	5.87	1.00	54.00	16.91	9.44	1.00	60.00	47.26	11.25	9.33	120.00
Min. credit duration (in month)	6.96	5.03	1.00	54.00	13.64	9.44	1.00	60.00	45.57	13.46	1.00	120.00
Max. credit duration (in month)	23.75	16.72	1.00	96.00	21.28	13.21	1.00	72.00	48.82	10.70	10.00	120.00
Interest rate of the granted credit	5.81	4.70	0.00	14.00	10.44	1.42	8.00	14.00	9.67	1.19	0.00	15.00
Mean interest rate of the granted credit	5.95	3.47	1.00	12.42	10.44	1.24	8.00	14.50	9.68	0.97	4.50	13.00
Min. interest rate of the granted credit	1.89	2.57	0.00	11.00	10.05	1.18	8.00	14.00	9.45	1.08	0.00	12.00
Max. interest rate of the granted credit	8.50	4.72	1.00	15.00	10.91	1.67	8.00	15.00	9.93	1.12	7.00	15.00

Table A1 – Continued from previous page

	Cluster 1 - FL-CF				Cluster 2 - EF				Cluster 3 - HV-LTF			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Dummy if frequency of credit is monthly (0= no, 1= yes)	0.99	0.08	0.00	1.00	0.95	0.21	0.00	1.00	0.02	0.13	0.00	1.00
Share of credits granted in January	0.04	0.12	0.00	1.00	0.09	0.24	0.00	1.00	0.01	0.07	0.00	1.00
Share of credits granted in February	0.04	0.13	0.00	1.00	0.08	0.22	0.00	1.00	0.04	0.17	0.00	1.00
Share of credits granted in March	0.07	0.17	0.00	1.00	0.10	0.25	0.00	1.00	0.13	0.31	0.00	1.00
Share of credits granted in April	0.13	0.23	0.00	1.00	0.08	0.23	0.00	1.00	0.26	0.39	0.00	1.00
Share of credits granted in May	0.09	0.17	0.00	1.00	0.10	0.24	0.00	1.00	0.35	0.43	0.00	1.00
Share of credits granted in June	0.12	0.25	0.00	1.00	0.08	0.23	0.00	1.00	0.12	0.30	0.00	1.00
Share of credits granted in July	0.20	0.27	0.00	1.00	0.09	0.23	0.00	1.00	0.04	0.17	0.00	1.00
Share of credits granted in August	0.08	0.13	0.00	1.00	0.09	0.23	0.00	1.00	0.02	0.14	0.00	1.00
Share of credits granted in September	0.11	0.23	0.00	1.00	0.06	0.15	0.00	1.00	0.01	0.05	0.00	1.00
Share of credits granted in October	0.04	0.10	0.00	1.00	0.09	0.23	0.00	1.00	0.01	0.09	0.00	1.00
Share of credits granted in November	0.04	0.14	0.00	1.00	0.07	0.21	0.00	1.00	0.01	0.07	0.00	1.00
Share of credits granted in December	0.04	0.13	0.00	1.00	0.07	0.20	0.00	1.00	0.01	0.06	0.00	1.00
Share of credits granted related to agriculture	0.01	0.09	0.00	1.00	0.15	0.35	0.00	1.00	0.05	0.16	0.00	1.00
Share of credits granted related to employment	0.52	0.36	0.00	1.00	0.83	0.37	0.00	1.00	0.01	0.11	0.00	1.00
Share of credits granted related to other reasons	0.00	0.06	0.00	1.00	0.02	0.12	0.00	1.00	0.94	0.19	0.00	1.00
Observations	777				1,244				1,314			

Table A2: Summary statistics on farmer repayment behaviour

	Cluster 1 - FL-CF				Cluster 2 - EF				Cluster 3 - HV-LTF			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Age of customer (continuous variable)	46.65	12.90	14.00	85.00	46.62	13.02	16.00	89.00	34.19	15.70	16.00	94.00
Gender of customer	0.87	0.34	0.00	1.00	0.88	0.33	0.00	1.00	1.00	0.06	0.00	1.00
Credit amount disbursed (continuous variable)	733771.39	1.38e+06	25,000.00	2.65e+07	2.67e+06	7.64e+06	25,000.00	1.00e+08	2.82e+06	4.59e+06	100000.00	1.00e+08
Credit amount disbursed (logarithmic value)	12.72	1.16	10.13	17.09	13.63	1.40	10.13	18.42	14.48	0.80	11.51	18.42
Interest rate of the granted credit	8.01	4.40	0.00	15.00	10.71	1.53	8.00	15.00	9.90	1.13	0.00	15.00
Interest rate (logarithmic value)	1.97	0.78	0.00	2.77	2.45	0.12	2.20	2.77	2.38	0.16	0.00	2.77
Credit duration (in month)	13.63	17.43	1.00	304.00	20.97	37.90	1.00	696.00	72.59	87.59	1.00	1,418.00
Dummy if credit granted in southern regions (0= no, 1= yes)	0.97	0.16	0.00	1.00	0.97	0.16	0.00	1.00	1.00	0.03	0.00	1.00
Number of due dates (in month, logarithmic value)	2.40	0.65	0.69	4.57	2.55	0.72	0.69	4.29	1.47	0.45	0.00	4.57
Overdue days (continuous variable)	8.28	10.67	-2.00	174.00	13.83	52.07	-3.00	946.00	36.35	99.75	-2.00	965.00
Dummy if credit is repaid in time (0= no, 1= yes)	0.35	0.48	0.00	1.00	0.39	0.49	0.00	1.00	0.18	0.39	0.00	1.00
Dummy if credit is repaid in time (0= no, 1= yes)	0.35	0.48	0.00	1.00	0.39	0.49	0.00	1.00	0.18	0.39	0.00	1.00
Dummy if repayment is delayed by more than 1 day (0= no, 1= yes)	0.55	0.50	0.00	1.00	0.50	0.50	0.00	1.00	0.43	0.50	0.00	1.00
Dummy if repayment is delayed by more than 15 days (0= no, 1= yes)	0.26	0.44	0.00	1.00	0.23	0.42	0.00	1.00	0.21	0.41	0.00	1.00
Dummy if repayment is delayed by more than 0 days (0= no, 1= yes)	0.00	0.07	0.00	1.00	0.03	0.18	0.00	1.00	0.15	0.36	0.00	1.00
Observations	4,132				3,146				2,191			