Data-driven & Theory-driven Science: Artificial Realities and Applications to Savings Groups

Rolando Manuel Gonzales Martinez

University of Agder

Doctoral Dissertations at the University of Agder 310

Gonzales Martínez, Rolando Manuel

Data-driven & Theory-driven Science: Artificial Realities and Applications to Savings Groups

Doctoral Dissertation for the Degree *Philosophiae Doctor (PhD)*
Specialization in International Business

Universitetet i Agder
Handelshøyskolen
2021

Doctoral Dissertations at the Universitetet i Agder 310
ISSN: 1504-9272
ISBN: 978-82-8427-014-2

Printed by 07 Media
Oslo

Preface

Funds and support were provided by the FAHU Foundation (Denmark) to analyze the *Savings Groups Information Exchange* (SAVIX) database. The SAVIX has information on more than 200000 savings groups in 52 countries worldwide. This dissertation discusses the epistemology, theoretical insights, and empirical findings obtained by applying both theory-driven and data-driven science to the SAVIX.

Acknowledgments

The support of FAHU Foundation and Annikka Berridge was fundamental to the project—without it, this work would be non-existent. Bert D'Espallier (KU Leuven) and Roy Mersland (Universitetet i Agder)—the PhD supervisors of this dissertation—gave their guidance throughout the project.

Rolando Gonzales Martínez
Kristiansand, Norway
January, 2021

Abstract

The scientific process is neither unique nor nomic—see Bradshaw et al. (1983) or the anarchist theory of knowledge of Feyerabend (1975). Two processes of scientific inquiry are theory-driven and data-driven science. This dissertation analyzes savings groups using theory-driven and data-driven methods. Simulated realities—based on data-driven theory—are used to understand the emerging dynamics of savings groups.

Savings groups are grassroots, community-based organizations composed of 15 to 30 members. These organizations—usually supported by international development agencies—have weekly meetings during a cycle of operations that typically lasts a year. In the groups, savings are kept in two funds: a fund for loans and a social welfare fund that covers life-cycle events.

The findings of **Papers A to D** in this dissertation provide new large-sample evidence about savings groups, their dynamics, and the factors affecting their financial performance. In practice, the results of **Paper A to D** shed light on the best policies to promote sustainable development with informal finance in a cost-effective way. A theory-driven approach indicates that the social fund in savings groups stimulates loan allocation among risk-sharing members, while implicitly covering idiosyncratic risks (**Paper A**). A data-driven approach based on Bayesian data-mining reveals that the macroeconomic environment and the facilitation model of development agencies have a strong influence on the profit-generating capacity of savings groups (**Paper B**). Machine-learning methods further show that business training is not the most frequent program implemented by development agencies, but it is in fact the most powerful intervention to encourage profits, particularly when a development agency stops working with a group and leaves a community (**Paper C**). Finally, the simulation of a village with artificial agents indicates that the businesses of savings groups can have higher profits due to the consolidation of social capital and the competitive advantage created through a process of homophily (**Paper D**).

Metatheoretically, the theory-driven and data-driven approaches of this dissertation—and the complementarity between these approaches—contribute to the epistemology of data-intensive science. The dissertation concludes that the gelstaltic and quasi-teleological explanations of the data-driven approach help to the formulation of theories through inductive and abductive reasoning.

Research papers

Research papers that are part of this dissertation:

Paper A: Gonzales Martinez, R., B. D'Espallier, R. Mersland (2020). Informal insurance and loan allocation in savings groups: The role of the welfare fund.

Paper B: Gonzales Martinez, R., B. D'Espallier, R. Mersland (2020). What drives profit generation in savings groups? Bayesian data-mining discoveries.

Paper C: Gonzales Martinez, R. (2019). Which social program supports sustainable grassroot finance? Machine-learning evidence. *International Journal of Sustainable Development & World Ecology*, 27(5): 1-7.

Paper D: Gonzales Martinez, R., B. D'Espallier, R. Mersland (2020). Bifurcations in business profitability: An agent-based simulation of homophily in self-financing groups. Forthcoming: *Journal of Business Research*.

Other research papers

Other research papers written during the PhD project:

Gonzales Martinez, R. (2018). The wage curve, once more with feeling: Bayesian model averaging of Heckit models. *Econometric Research in Finance*. 3(2): 79-92.

Gonzales Martinez, R., A. Rojas-Hosse (2019). Inflation shocks and income inequality: An analysis with genetic algorithms and Bayesian quantile regressions. *African Journal of Economic and Management Studies*. 10(2): 226-240.

Gonzales Martinez, R., G. Aguilera-Lizarazu, A. Rojas-Hosse, P. Aranda (2019). The interaction effect of gender and ethnicity in loan approval: A Bayesian estimation with data from a laboratory field experiment. *Review of Development Economics*. 24(3): 726-749.

Samonte, F., M. T. De Guzman, R. Gonzales Martinez (2020). A population level analysis of mental health and non-communicable disease (NCD) in the Philippines using predictive modelling analysis. *International Journal of Psychosocial Rehabilitation*. 24(6): 12149 - 12158.

Anand, P., S. Saxena, R. Gonzales Martinez, H. Dang (2020). Can Women's Self-help Groups Contribute to Sustainable Development? Evidence of Capability Changes from Northern India. *Journal of Human Development and Capabilities*. 21(2): 137-160.

Anand, P., H. Allen, R. Ferrer, N. Gold; R. Gonzales Martinez, E. Kontopantelis, M. Krause, F. Verguns (2020). Work Related and Personal Predictors of COVID-19 transmission. IZA discussion paper 13493.

Ravelli, E., R. Gonzales Martinez. Environmental risk factors of airborne viral transmission: Humidity, influenza and SARS-CoV-2 in the Netherlands. August 2020. Preprint: medRxiv 2020.08.18.20177444

Alkire, S., P. Conceição, C. Calderón, J. Dirksen, M. Evans, R. Gonzales Martinez, J. Hall, A. Jahic, U. Kanagaratnam, M. Kivilo, M. Kovacevic, F. Kovesdi, C. Mitchell, R. Nogales, A. Ortubia, M. Pinilla-Roncancio, N. Quinn, C. Rivera, S. Scharlin-Pettee, N. Suppa (2020). Charting pathways out of multidimensional poverty: Achieving the SDGs. Technical report: Oxford Poverty and Human Development Initiative (OPHI) at the University of Oxford.

Other documents written during the PhD project:

Gonzales Martinez, R. No access to formal banking? How nano-finance is unlocking the power of communities around the world. SABE-TFI Impact Essay Award 2019: First place for relatable focus on financial wellbeing.

Contents

Chapter 1

Introduction

Theory-driven analysis and data-driven analysis are two prominent processes of scientific inquiry. In the theory-driven approach, a hypothesis is proposed and tested based on an expected pre-defined theoretical cause-effect relationship. In the data-driven approach, data are first analyzed to discover patterns, which are the basis for the formulation of theories (Kelling et al., 2009).

Data-driven methods are important in many scientific applications. The large Hadron Collider of the European Organization for Nuclear Research (CERN) applies data-driven algorithms to reconstruct particle tracks left in silicon detectors (Chen & Zhang, 2014). In bio-medicine, the analysis of genome sequences starts with the mining of biologically relevant patterns and afterwards hypotheses are generated for experimental validation (Blake & Bult, 2006). In development studies, Holloway et al. (2018) apply machine learning for monitoring the United Nation's Sustainable Development Goals.

This dissertation applies methods of theory-driven and data-driven science to large-sample information on savings groups. Savings groups are community-based organizations that provide informal financial services to 30 million people around the world, mostly in Africa (Seel, 2018). These associations are formed, owned and managed by 15 to 30 members—usually women—who meet weekly during a cycle of operations that typically lasts a year. In the groups, savings are kept in two funds: a fund to provide loans with fixed interest rates to selected members, and a social fund for members' welfare.

The activities of savings groups follow a cycle system: at the beginning of a cycle, groups are formed, leaders are elected, and members agree about the duration of the cycle, as well as on issues like the contribution of each member to the common fund, the frequency of meetings, fines, and the procedures for requesting a loan. Saving and borrowing take place during the cycle. At the end of the cycle, all funds gained in a group—accumulated savings and retrieved interest—are distributed among the members according to their savings balances. In the next cycle, the entire process is repeated. Due to the meaningful role of savings groups for poverty alleviation, major donors and development agencies—e.g., the Inter-American Development Bank, the Bill & Melinda Gates Foundation, CARE, and Oxfam International, to name a few—currently work with these groups as a platform to promote sustainable development and financial inclusion through social programs.

In this dissertation, **Paper A** applies a theory-driven approach to identify the externalities caused by the implicit insurance mechanisms of the social fund in savings groups. The theoretical mechanism proposed in **Paper A** suggests that behavioral changes arise from the higher affective and cognitive trust stimulated by the existence of a welfare fund, which creates incentives for increasing the number and amount of loans in savings groups. This hypothesis is tested with panel regressions and quasi-experimental methods. The findings indicate that the welfare fund goes beyond its social role and acts as an informal risk-sharing technology to cope with idiosyncratic shocks, which encourages members to take more risks and larger loans.

Paper B and **Paper C** apply data-driven methods to identify the covariates influencing the profit-generating capacity of savings groups. In **Paper B**, Bayesian data-mining methods are applied to a database of more than 200,000 savings groups worldwide in order to identify which micro, meso, and macro factors are related to profit generation. The results show that external factors are more important than internal group dynamics for profit generation. **Paper C** applies text-mining and machine-learning methods to identify which social programs are good predictors of financial returns in savings groups. The results indicate that education, income-generating activities, and health programs are the most frequent social programs provided by development agencies to savings groups. Business training is not the most frequent intervention, but it is in fact the most important social program to encourage financial sustainability, particularly after a development agency stops working with a group and leaves the community.

Finally, in **Paper D**, the theory about the dynamics of savings groups and the empirical findings of **Papers B** and **C** are synthesized with complex algorithms. These algorithms simulate an artificial village with savings groups formed by artificial agents. **Paper D** compares the profitability of businesses started by members of savings groups to the profitability of businesses financed by commercial loans. The results indicate that—besides the self-generation of debt capital—businesses of members of savings groups can have higher returns due to the consolidation of social capital and the competitive advantage created through a dual process of homophily. Higher quotas of savings boost profits, but only up to a threshold, after which a bifurcation pattern—typical of complexity dynamics—emerges.

The findings of **Papers A to D** contribute to the literature on informal finance at the bottom of the wealth pyramid by providing new theory-driven and data-driven evidence about the dynamics of savings groups and the factors affecting the performance of these associations. From an epistemological point of view, the dissertation concludes that data-driven and theory-driven science complement each other in a circular fashion: an extension of Eisenhardt's cyclical model is proposed to illustrate how data-driven methods contribute to theorizing with gestaltic and quasi-teleological explanations that arise from induction and abductive reasoning.

The next chapter provides a brief background of theory-driven and data-driven science. Chapter 3 summarizes the applications of theory- and data-driven methods to the information on savings groups. Chapter 4 concludes.

Chapter 2

Theory-driven and data-driven science

The ontological and epistemological underpinnings of a scientific approach depend on the tradition and interest of researchers (Elragal & Klischewski, 2017). A theory-driven approach is based on the falsifiability of hypothesis, while in data science the data-driven approach discovers patterns in data without a pre-defined hypothesis[1].

The falsifiability of the hypothesis in a theory-driven approach is a rejection-based solution to the problem of induction (Popper, 1963). Theory-driven science—framed under hypothetico-deductivism—formally starts by using theory to produce a hypothesis, which is afterwards rejected or not rejected based on data evidence. Even ad hoc hyphoteses with weak theoretical content—which were regarded as undesirable by Popper due to circularity and lack of testable empirical content (Bamford, 1993)—are considered necessary for the generation of new theories according to Lakatos et al. (1980). In **Paper A** of this dissertation, a theory-driven approach is used to formulate a hypothesis about the financial externalities that arise from the trust-based mechanisms of social funds in savings groups.

The theory-driven approach is not free of limitations. In some situations, instead of a rigorous upfront theory, only a narrative selection of related literature—a 'literature review'—is presented and linked to a hypothesis (Markus, 2014). As theories are not always correct or complete, theory-driven science tends to be applied as a type of theory-light research. Even with the development of a formal theory, a problem with a theory-driven approach is that a hypothesis can restrict the propositions deduced from phenomena, thus limiting the amount of knowledge that can be acquired from a rigorous analysis (Harper, 2011).

A data-driven renaissance emerged as an alternative to the theory-driven approach. The increasing availability of large amounts of information, higher computational capacity, and the development of novel statistical methods encouraged the use of computationally intensive data-driven methods. Algorithms 'learn' from data in these methods (Kodratoff, 2001), as in data-mining, big-data analytics or machine learning—see Alpaydin (2016), Witten et al. (2016) or Parish and Duraisamy (2016).

Mazzocchi (2015) traces the epistemological roots of the data-driven approach to the

[1]There is an ambiguity in the definition of data science, and a debate exists about whether data science is a new discipline different from statistics; see for example Donoho (2017). Synthesizing the definitions provided by Cao (2017, pp.43:8), data science can be considered the interdisciplinary science that transforms data into insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology, with the ultimate goal of obtaining data products like e.g. a discovery, prediction or model. In this thesis, data-driven science is used as synonym of data science, while the data-driven approach is conceptualized as the analytical endeavor applied in data science to obtain data products.

empiricist ideas of Francis Bacon and Isaac Newton. In his *Novum Organum*, Bacon (1620) makes a distinction between men of experiments (ants) and men of dogmas (spiders), but highlights the existence of a middle course (bees). In his interpretation of the ideas of Bacon, Boden (2004) conceptualizes science as data-driven because scientists look for regularities and anomalies in observational data. Harper (2011, Chapter 1, VII) argues also that Newton's scientific method goes beyond the hypothetico-deductive model and includes propositions derived from induction. Moreover, Harper (2011) interprets Newton's *hypotheses non fingo* passage as one of the first signs of the diminishing role of conjectured hypothesis in experimental philosophy.

Currently, data-driven science is experimenting a revitalization in the form of technological empiricism. The use of data-driven methods is advocated as a fourth paradigm in science, not in the sense provided by Kuhn (1970)—who argues that a paradigm shift occurs when the dominant mode of science cannot account for particular phenomena—but rather in the sense of Hey et al. (2009), who claim that scientific transitions are founded on advances in forms of data and the development of new analytic methods and technologies to analyze data.

The classification into either data-driven or theory-driven science is not sharp and does not imply that the data-driven approach is lacking in theory or that data-driven steps are not part of the theory-driven approach (Niemeijer, 2002). Assume for example that data science can be conceptualized as the process of using complex algorithms for the *compression* of a large data set $\mathbf{D}$ into a reduced statistical object δ (Box 1)[2]. A more complex program—in the sense of conditional Kolmogorov algorithmics—will be needed to construct and calculate δ in a data-driven approach compared to a theory-driven approach, but the compression process is not necessarily theory-free. Moreover, the explanations that arise from data-driven science can lead to the formulation of theories through inductive or abductive reasoning.

Similarly, in a Bayesian epistemic framework, the tension about the role of theory in the data-driven approach may be baseless: for example, let $\mathcal{L}_{\mathcal{S}}(\mathbf{D}|\Theta)$ be a likelihood function and $\pi(\Theta)$ be the prior probability density function of Θ (defined in Box 1), then by Bayes theorem,

$$\pi(\Theta|\mathbf{D},\mathcal{S}) \propto \pi(\Theta)\,\mathcal{L}_{\mathcal{S}}(\mathbf{D}|\Theta),$$

where $\pi(\Theta|\mathbf{D},\mathcal{S})$ is the posterior density function of Θ conditional on any theoretical claim or previous empirical results about Θ, besides the data evidence $\mathbf{D}$, $\mathcal{S}$ and the set of human and/or machine-made decisions.

[2]In the case of Netflix, for example, the large data set $\mathbf{D}$ with information about viewers' preferences can be analyzed with data-driven methods to produce a compressed result in terms of δ-recommendations of videos (see Bell & Koren, 2007). Other δ representations based on $\mathbf{D}$ can be point estimates of the average rating of viewers, or the joint multivariate distribution of age and income of Netflix subscribers.

The corollary of having more complex algorithms in data science is that some type of automation of the scientific process is implemented for cleaning and analyzing the data. Formally, let $\mathbb{H} = \bigcup_{h \in H} \mathcal{P}_h$ be the set of $h = 1, 2, ..., H$ $\mathcal{P}$-procedures employed to obtain δ in a theory-driven approach, and let $\mathbb{D} = \bigcup_{d \in D} \mathcal{P}_d$ be the corresponding set of the data-driven approach. Denote the $\mathcal{S}$-machine-made decisions in the supersets $\mathbb{H}$ and $\mathbb{D}$ as $\mathbb{H}_m \subset \mathbb{H}$ and $\mathbb{D}_m \subseteq \mathbb{D}$, respectively. If $\mathcal{K}_{\mathbb{D}}(\delta|\mathbf{D}) \gg \mathcal{K}_{\mathbb{H}}(\delta|\mathbf{D})$, then $\bar{\bar{\mathbb{D}}}_m > \bar{\bar{\mathbb{H}}}_m$.

A similar notion to algorithmic complexity is algorithmic probability, which was proposed by Solomonoff (1964) as a formal theory of inductive inference. Roughly, given a set of hypothesis h and the data set $\mathbf{D}$, a way to decide which hypotheses has the higher conditional probability $\mathbb{P}(h|\mathbf{D})$ is through the Bayes rule $\mathbb{P}(h|\mathbf{D}) \propto \mathbb{P}(\mathbf{D}|h)\mathbb{P}(h)$. Algorithmic probability uses two philosophical principles to suggest a value for the prior probability $\mathbb{P}(h)$: keep all hypotheses that are consistent with the data (Epicureus) and choose the simplest one (Occam's razor).

Box 1: Data-driven science and algorithmic complexity

Proposition 1: Data-driven science as data compression. *Data-driven science can be conceptualized as the process of creating a* δ*-statistical representation of a data set* $\mathbf{D}$ *in the presence of Kolmogorov algorithmic complexity* $\mathcal{K}(\delta|\mathbf{D})$.

In Proposition 1, data science aims to represent data with a compressed statistical description through data-driven methods (Manin, 2013). This proposition is based on the definitions of algorithmic complexity (Lemma 1) and the δ-representation of a data set (Lemma 2).

Lemma 1: Algorithmic complexity. *Let* δ *be a description of a data matrix* $\mathbf{D}$*, and* p *a program that produces* δ *conditional on* $\mathbf{D}$ *when run on a universal Turing machine* T*, for* $|p|$ *the length of* p*. The algorithmic complexity of* δ *can be defined with the conditional Kolmogorov complexity* $\mathcal{K}(\delta|\mathbf{D})$*:*

$$\mathcal{K}(\delta|\mathbf{D}) = \min\{|p|, p : T(p, \mathbf{D}) = \delta\}$$

Lemma 1 states that a computer program p is applied to a database $\mathbf{D}$ with the aim of producing a compressed description of $\mathbf{D}$ in terms of a δ statistical object. Intuitively, an algorithm to produce δ can be considered simple if it can be described in a few lines of code, and it will be complex if there is no such short description. Under a data-driven approach, a more complex program—in the sense of Kolmogorov complexity—is needed to construct and calculate a δ representation object.

Lemma 2: δ***-representation.*** *Let* $\mathbf{D} \ni \{\mathbf{Y}, \mathbf{X}\}$ *be a data matrix of dimension* $n \times (k+1)$*, where* $\mathbf{X}$ *is a partition of explicans for* $\mathbf{Y}$*. A probabilistic model to obtain a* δ*-representation of* $\mathbf{D}$ *is:*

$$\begin{cases} \pi(\mathbf{Y}|\mathbf{X}, \hat{\Theta}, \mathcal{S}) \overset{\rightarrow}{\equiv} (1-\gamma)\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right) + \gamma\epsilon \\ \epsilon \sim \mathrm{IID}(0, \sigma^2_\epsilon), \\ \mathbf{X} = [x_{n,1} \quad x_{n,2} \quad \cdots \quad x_{n,k}], n \in \mathbb{Z}_+ \ k \in \mathbb{Z}_+ \\ \mathbf{D} \ni \{\mathbf{Y}, \mathbf{X}\}, \ \Theta \in \mathbb{R}^k, \ \Omega \in \mathbb{R}_+^{k\times k} \ \gamma \in \mathbb{R}_{0,1} \end{cases}$$

Lemma 2 suggests an additive mixture of two distributions as the explicit form of δ. The form of δ in Lemma 2 is convenient but not unique: the object δ can be, for example, a point estimate of interest or a distribution function estimated with an algorithm applied to the $\mathbf{D}$ data set. (The δ object is what Cao (2017, pp.43:8) more generally calls the 'data product' of data science.) In the conditional probability density function $\pi(\mathbf{Y}|\mathbf{X}, \hat{\Theta}, \mathcal{S})$, $\mathbf{Y}$ depends on observed k-potential *explicans* contained in $\mathbf{X}$ (equation 2), weighted by the parameters in Θ—which assign the importance of each $\{x_{n,1}, \ldots, x_{n,2}\} \ni \mathbf{X}$ to the *explicandum* $\mathbf{Y}$—and a set $\mathcal{S}$ of human and/or machine-made decisions about the model, which are based on additional information not available in the data. This additional information, if irrelevant, will cancel out mathematically—see the last part of Chapter 3 and Chapter 4 in Jaynes (2003).

The conditional density of $\mathbf{Y}$ is a γ-convex combination of an independent and identically distributed (IID) error term ϵ and a density $\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)$. In the density $\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)$, $\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right]$ is the expectation of a function $f\left(\mathbf{X}, \Theta, \mathcal{S}\right)$ and Ω is a $k \times k$ (positive-definite) variance-covariance matrix.

The scalar $\gamma \in \mathbb{R}_{0,1}$ weights the error term ϵ. The value of γ measures the degree to which a precise prediction of $\mathbf{Y}$ can be made with $\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)$: the predictability of a phenomenon decreases as $\gamma \to 1$ and increases as $\gamma \to 0$. The limit $\gamma = 0$ produces a degenerate distribution for $\mathbf{Y}$ and is a special case called Laplace's Demon in philosophy. (For example, the returns in stock markets or the evolution of El Niño-Southern Oscillation during the spring predictability barrier are phenomena where $\gamma \to 1$, while $\gamma \to 0$ in e.g. eclipses.)

In the model, $\mathbf{Y}$ can be continuous or discrete ($\mathbf{Y} \in \mathbb{R} \vee \mathbf{Y} \in \mathbb{Z}$, corresponding to regression or classification, respectively). More importantly, the symbol $\overset{\rightarrow}{\equiv}$ highlights that data science is a process of orthogonal decomposition of the observational data contained in $\mathbf{Y}$ into an element that can be explained by the density $\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)$ and a residual ϵ:

$$\underbrace{\pi\left(\mathbf{Y}|\mathbf{X}, \hat{\Theta}, \mathcal{S}\right)}_{\text{explicandum}} \overset{\rightarrow}{\equiv} (1-\gamma)\underbrace{\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)}_{\text{explicans}} + \underbrace{\gamma\epsilon}_{\text{residual}}$$

The previous equation should not be confused with

$$\underbrace{\pi\left(\mathbf{Y}|\mathbf{X}, \hat{\Theta}, \mathcal{S}\right)}_{\text{explanandum}} \overset{\leftarrow}{\equiv} (1-\gamma)\underbrace{\pi\left(\mathbb{E}\left[f\left(\mathbf{X}, \Theta, \mathcal{S}\right)\right], \Omega\right)}_{\text{explanans}} + \underbrace{\gamma\epsilon}_{\text{residual}}$$

where the symbol $\overset{\leftarrow}{\equiv}$ indicates *causality* from $\mathbf{X}$ to $\mathbf{Y}$ in randomized experiments. A deliberate distinction is made in the use of the terms explicandum/explicans *vis a vis* explanandum/explanans: just as in Hempel and Oppenheim (1948), the terms *explicandum/explicans* in the observational equation express the analysis that looks for an explication of meaning, whereas in the causal equation the *explanandum* is a logical consequence of the *explanans*. The notation and terminology of the observational decomposition is congruent with Carnap's principle of total evidence (Carnap, 1945), i.e. the recommendation to use all available evidence when estimating a probability (Good, 1967).

The correlational structures of Θ provide a starting point for understanding the relation between phenomena and theory-building (Leonelli, 2014) . Recent advances in orthogonal machine-learning modify the correlational estimators Θ to account for counterfactual causality—for some examples see Chernozhukov et al. (2017), Mackey et al. (2017), Semenova (2018), Kreif and DiazOrdaz (2019) or the developments in orthogonal random forests of Oprescu et al. (2018). With these modified estimators, machine-learning algorithms are used to create a balance among treated and control groups, as well as to estimate the conditional expectations of the outcome or to select variables when there is a high number of covariates. Data science, then, not only provides correlations, but can also identify counterfactual causality in the sense of Lewis (1973).

The prior $\pi(\Theta)$ provides an explicit opportunity to systematically include theory or any other type of non-quantitative information into the statistical model. In a theory-light approach, $\pi(\Theta)$ will be chosen to be uninformative by using e.g. a Jeffrey's prior,

$$\pi(\Theta) \propto \sqrt{-\mathbb{E}_{\mathbf{X}|\Theta}\left[\left(\frac{\partial \log f(\mathbf{X}|\Theta)}{\partial\Theta}\right)'\left(\frac{\partial \log f(\mathbf{X}|\Theta)}{\partial\Theta}\right)\right]},$$

or a flat/uniform prior where the role of theory diminishes as the sample size n increases (Bauwens et al., 2000). At the other extreme, substantive knowledge—be it theoretical, based on expert opinion, and/or previous empirical evidence—can be deliberately included in data-intensive science by elicitating $\pi(\Theta)$ on the basis of e.g. Leamer (1983) hierarchy of axioms to conventions; see Gill (2015). It follows that, under a Bayesian paradigm, there is not a sharp distinction between theory-driven and data-driven science in terms of the inclusion or exclusion of theory, but rather a *continuum* of possibilities depending on the degree of elicitation of the prior $\pi(\Theta)$.

In this dissertation, theory is not neglected but only minimized during the process of data compression by data-driven methods. Data-driven methods reduce the need for strong domain theories as starting points for a scientific analysis, hence creating new epistemic models that supplement and extend the traditional scientific method (Janowicz et al., 2015). The role of theory can be minimized in data-driven science through a temporary abstention from theoretical concepts—this is, through what is refered below as a voluntary *epochē*. This is the approach taken in **Paper B** and **Paper C**.

The *epochē* (ἐποχή) is defined by Sextus Empiricus as the standstill of the intellect owing to which nothing is postulated nor rejected (Empiricus, 2000). The *epochē* can be understood from the point of view of the philosophy of skepticism or the phenomenological conceptualization of Husserl. In skepticism, *epochē* is a part of a larger three-stage exercise: equipollence, *epochē*, and *ataraxia* (Massie, 2013). In the modern conceptualization provided by Husserl, *epochē* is a refraining from judgment, a suspension of belief which involves no skeptical doubt, but rather makes a temporary 'bracketing' (*Einklammerung*) of the role of theory (Drummond, 2007).

In data science, the *Einklammerung* provisionally allocates theoretical concepts and preconceptions into a parenthesis, temporarily suspending the beliefs of a scientist. Based on Husserl's two moments of phenomenological reduction—the *epochē* and the reduction proper (Luft, 2004)—, data-driven science can be seen as an abstention (*enthaltung*), a transitory 'unplugging' (*ausschaltung*) of the positing about the world and the reality of what is experienced, which implies putting temporarily out of action (*außer aktion zu setzen*) or out of play (*außer spiel zu setzen*) theoretical judgments (Moran & Cohen, 2012). The second moment, the reduction proper—the transcendental insight that a theory is just an acceptance about the nature of a phenomenon and not an absolute—is fundamental for a proper balance between the role of theory and the weight of data evidence.

When the temporary 'bracketing' of theory is lifted, inductive and abductive reasoning allow the data scientist to make sense of the emerging patterns obtained with data-driven methods. Inductive learning implies using samples of data subsets—of inputs and an output target—to estimate a function able to generalize to new samples in the future (Domingos, 2015). This is the case, for example, of deep-learning algorithms (see Deng, Yu, et al., 2014), which can be justified in terms of the universal approximation theorem (Cybenko, 1989; Hornik, 1991; Lu et al., 2017) and are applied in computer vision and automatic speech recognition.

The type of induction that arises from the application of data-driven methods conforms better with eliminative induction and enumerative induction, than with demonstrative induction. Data-driven methods—that are based on using multiple data subsamples for generalization—do not follow the one-case generalization of the mixed hypothetical syllogism *modus ponendo ponens* of demonstrative induction (Broad, 1968), i.e. given a sample $x_s \in \mathbf{X}$, if a result is in (at least one sample) $x_s \subset \mathbf{X}$ is $\hat{\Theta}_s$ then it cannot be said that $\hat{\Theta}_1 = \hat{\Theta}_2 = \cdots = \hat{\Theta}_s$ in all x_s, for $s > 1$. By contrast—and in line with the first induction theorem of Good (1975, pp. 62) that provides an account of enumerative induction—, in data-driven science, instances estimated from multiple samples increase the probability of other potential instances in the limit when the number of subsamples s and the sample size n tend to infinity:

$$\lim_{\substack{s\to\infty^+\\ n\to\infty^+}} \mathbb{P}\left(\hat{\Theta}_{s+1}\hat{\Theta}_{s+2}...\hat{\Theta}_{s+n}|\hat{\Theta}_1\hat{\Theta}_2...\hat{\Theta}_s\right) \to 1$$

where $\hat{\Theta}_1\hat{\Theta}_2$ is the conjunction of $\hat{\Theta}_1$ and $\hat{\Theta}_2$, and so on; see Zabell (2011) and the proof of the theorem in Huzurbazar (1955), who shows that inductive inference can approach certainty, but it cannot reach it. Also, Hawthorne (1993) notes that Bayesian induction is a form of eliminative induction, since in a Bayesian data-driven approach posterior probabilities and Bayes factors can be used to discard instances with low supporting evidence. This is the approach taken in the Bayesian data-mining algorithm of **Paper B**.

Abduction also plays a role in data science when data-driven methods produce unanticipated empirical results. In the classic definition—provided by Peirce (1960)—, abduction is the process of forming an explanatory hypothesis. Piekkari and Welch (2018) conceptualize abduction as a process of reasoning aimed to provide a theoretical explanation for an empirical puzzle triggered by an observation that challenges existing theoretical frameworks and preconceptions—an unmet expectation that creates a discontinuity and asks for a theory to make the observation meaningful (Van Maanen et al., 2007).

Based on the classification of Magnani (2011), the kind of abduction applied in data-driven science can be considered both selective and creative, as data science can lead to choosing an optimal explanation from a myriad of possible explanations (selective abduction), but it can also introduce new theoretical models or concepts based on unexpected findings (creative abduction). The difference between inductive and abductive data-driven methods is that abductive data science starts the research process with more theoretical content than inductive data-driven science, in order to allow for a contradiction between the expected findings and the surprising, anomalous empirical results (Tavory & Timmermans, 2014).

The inductive and abductive reasoning processes of data-driven science complement the deductive approach of theory-driven science. Kitchin (2014) notes that data-driven science incorporates induction into the research design to identify potential questions—hypotheses—worthy of further examination and testing, before a deductive approach is employed. In contrast to the experimental deductive design, data-driven science accommodates minimal theoretical concepts to generate hypotheses and insights 'born from the data' rather than 'born from the theory' (Kelling et al., 2009)[3].

[3]While experimentation is seen as a gold standard in scientific practice, Hendry and Doornik (2014) highlight that experimentation is neither necessary (astronomy) nor sufficient (alchemy) to define science. Serendipity can be a source of scientific discovery, if a meaning is properly assigned to the occurrence of random events, as in Fleming's discovery of penicillin or Penzias and Wilson's discovery of cosmic micro-wave radiation.

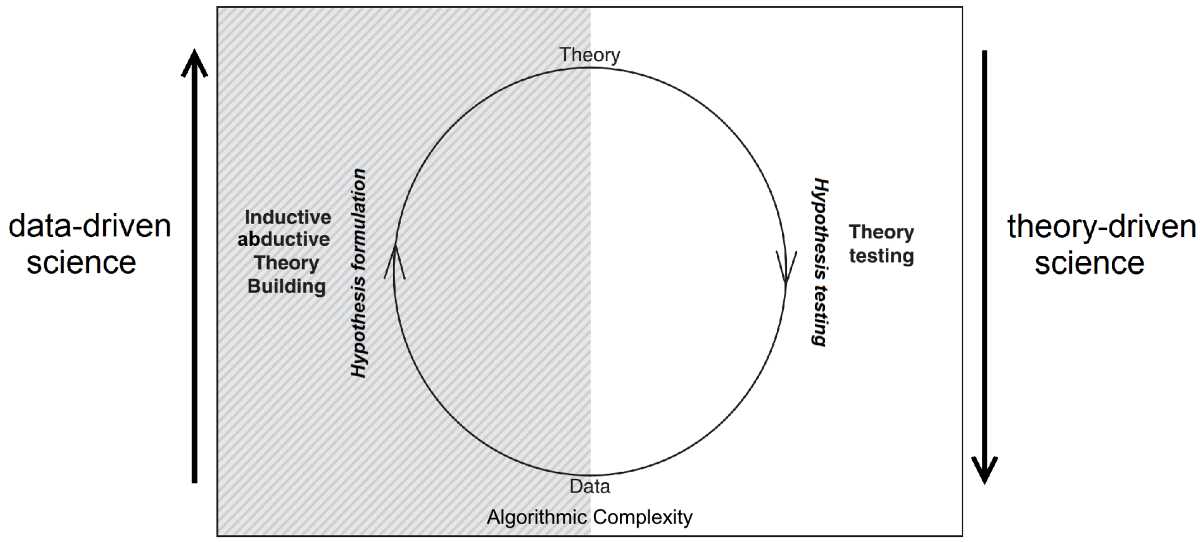


Figure 2.1: Illustration of the cyclical process of research with theory-driven and data-driven science. The figure is an extension of the Eisenhardt (1989) positivist view about the interaction between inductive and deductive reasoning (originally discussed and illustrated in Piekkari and Welch, 2011, page 8). Source: modified from Piekkari and Welch (2018).

As Kitchin (2014), Eisenhardt and Graebner (2007, p.25) argue that inductive and deductive logic are two halves of a cyclical process in which induction helps to formulate new theory from the data and deduction tests theory. Eisenhardt (1989) and Eisenhardt and Graebner (2007) develop a cyclical model of theorizing with case studies based on this complementarity of induction and deduction. If Eisenhardt's model is extended to replace case studies with data-driven methods (Figure 2.1), data-intensive science can be conceived as an additional approach to the development of testable hypotheses and theory through inductive or abductive reasoning. Similarly to case studies, the data-driven approach generates explanations—through induction and abduction—from empirical constructs—e.g. principal components or factor analysis—, from empirical measures—as those that emerged from the patterns identified with machine-learning—, and from testable propositions consistent with deductive research.

The types of explanations that arise from data-driven science can be framed under the typology of explanations provided by Evered (1976), who draws a distinction between causal explanations, gestaltic, and teleological explanations, according to the relationship between the *explicans* and the *explicandum* (Figure 2.2). In the framework of Evered, through induction and abduction, data-driven methods provide gestaltic and quasi-teleological explanations for the patterns discovered in data.

In gestaltic explanations, events are understood from the pattern of their relationships with other events, and the meaning attached to the events depends upon the patterns and interconnections discovered through data-driven methods. In **Paper B** and **Paper C**, for example, the emerging relationships between variables lead to the discovery of the variables affecting the profit-generating capacity of savings groups, without the explicit formulation of a theory ex ante. **Paper D** in turn provides quasi-teleological explanations, because, in the complex algorithms of **Paper D**, the artificial agents make decisions *in order to* achieve specific social and economic goals, just as in the anthropological sciences.

In the next chapter, applications of theory-driven and data-driven science to observed and artificial data of savings groups show the similarities, differences and complementarities

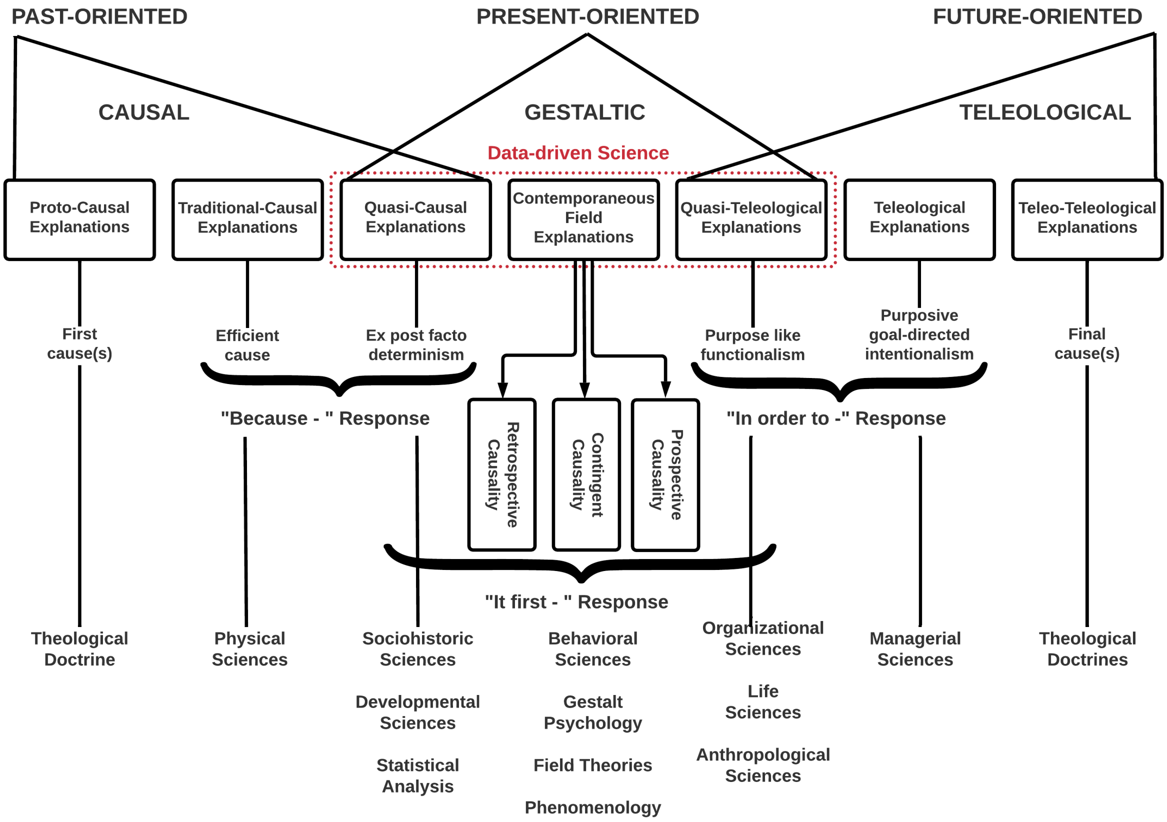


Figure 2.2: Typology of explanations based on the relationship of the explicans and the explicanda, according to Evered (1976). Data-driven science provides gestaltic explanations. Quasi-teleological explanations arise from data-intensive methods based on complex algorithms. Source: modified from Evered (1976, Fig. 5, p. 273).

between both approaches. **Paper A** applies a theory-driven approach based on theory formulation and hypothesis testing using deductive methods. **Papers B** and **C** apply data-driven methods to produce gestaltic explanations through abductive reasoning, on the basis of eliminative induction (**Paper B**) and enumerative induction (**Paper C**). The quasi-teleological explanations of **Paper D**—obtained with data-driven theory—complement the empirical findings of **Papers A, B**, and **C**.

Chapter 3

Empirical applications and simulations

This chapter summarizes **Papers A, B, C**, and **D**, which apply theory-driven and data-driven science to the information on savings groups. In savings groups, low-income individuals without access to formal financial services create a group and start to accumulate their savings into funds, which they later use to provide themselves with loans and insurance. See details in **Box 2**.

The information on savings groups used in the empirical applications was obtained from the SAVIX database. In 2009, the Bill & Melinda Gates Foundation initiated three large-scale savings group projects and commissioned Village Savings & Loan (VSL) Associates to build a web-based database to compare the performance of these groups (Allen & Panetta, 2010). This initiative led to the creation of an online reporting system, the Savings Groups Management Information System (MIS), which is synchronized with the *Savings Groups Information Exchange* (SAVIX) database. The MIS enables users to collect and store group-level data on over 200000 savings groups in different regions of the developing world.

The SAVIX database is an unbalanced longitudinal (panel) data set that has quarterly observations from 2010 to 2017 as the time index and group-level variables as cross-section identifiers. The database has a total of 233,180 savings groups in 52 countries around the world, with a high concentration in Africa. The SAVIX includes information about group-level characteristics, financial variables, and covariates related to the activities of development agencies working with the groups.

The first empirical application to the SAVIX data in this dissertation is based on a theory-driven approach: **Paper A** discusses the externalities caused by the informal insurance provided by savings groups. In the second and third applications (**Paper B** and **Paper C**), data-driven methods discover the covariates influencing the profit-generating capacity of savings groups. The chapter concludes with an application of a theory-driven but data intensive approach based on complex algorithms that simulate an artificial village where savings groups thrive (**Paper D**).

Box 2: Savings Groups

Savings groups are community-based providers of informal financial services. These associations are formed by 15 to 30 members——usually women—— who meet weekly during a cycle of operations that typically lasts 9 to 12 months. Savings group members deposit their mandatory savings into a metal box during the meetings and afterwards loans are allocated to selected members with the money from the box. Interest rates are fixed by the group. Some groups also own some property like a calculator or a small hut where they hold their meetings. During the cycle, groups keep part of the collected savings as cash in a box due to lack of demand for loans or groups' need to accumulate cash before the payout at the end of the cycle. At the end of the cycle, all the collected funds and earnings are distributed among the members according to their savings balances.

The average size of savings groups was explained by Bisrat et al. (2012) as follows: while a higher number of members is needed to accumulate a large sum of money over a cycle, too many participants leads to administrative problems as well as a long cycle time, thus creating an incentive to keep the number of participants around 20 members. Abbink et al. (2006) explain that the higher female composition in savings groups is a consequence of NGOs targeting women partly because they consider women's empowerment as a goal but also because women are often seen as more reliable borrowers. Rasmussen (2012) relates the gender composition to women's economic resilience, since savings enable to cope with income shocks and tide over unforeseen emergencies such as illness or loss of employment (Ghosh & Vinod, 2017). Anderson and Baland (2002) suggest that intra-household conflict can explain the high incidence of female participation in savings groups. Guha and Gupta (2005) add that targeting women over men in rural areas is based on the premise that women make a higher contribution to family welfare, since women give priority to spending their earnings on their children.

Historically, savings groups can be traced back to primal models of mutual financial association, such as the *Hui* in China during the Tang dynasty (618-906 CE) or the *Shê*, where a man who was in need of money (*Shê-chu*) invited others to cooperate with him and when the requisite number had been secured, the members (*Shê-yu*) assembled and fixed the order in which each would have the use of the common fund. Similar associations are the *Kou*, a group of savings and loan associations in Japan in 1275 CE (McKeever, 2009), the *Modjokuto* in Indonesia (Geertz, 1962), and the *Kameti* in India. In West Africa, the main indigenous savings and lending associations are called *esusu*—or its variants, like the *SuSu* clubs in Ghana (Anku-Tsede, 2013)—, while in Central and East Africa they are referred to as *ikilemba* (Ardener, 1964).

Modern savings groups evolved from two indigenous grassroots financial systems: Rotating Savings and Credit Associations (ROSCAs) and Accumulating Savings and Credit Associations (ASCAs). In ROSCAs, the pooled savings are distributed to the members in rotation, until each member has received a share. In a random ROSCA the savings are set aside based on a random drawing of lots, while in a bidding ROSCA the participants bid competitively for the pool of savings, which is allocated to the highest bidder (Handa & Kirton, 1999). In ASCAs the savings are not instantly redistributed but rather accumulate in a fund in order to make loans with a fixed maturity. At the end of the cycle, redistribution takes place and a new cycle starts.

Facilitated savings groups are promoted by and receive training from development agencies or the government. International NGOs—in collaboration with their local partners—have designed their own group models based on the principles of ROSCAs and ASCAs. According to le Polain et al. (2018), the best-known group models are the Village Savings and Loan Associations (VSLAs) initiated by CARE International and the Savings and Internal Lending Communities (SILCs) promoted by Catholic Relief Services. CARE created the VSLA models based on the ASCA methodology and the indigenous model of Mali's local *tontines* (Karlan et al., 2017). The VSLA model evolved from an approach designed only for impoverished and uneducated rural women into a model for both literate and illiterate men and women living in rural areas and urban slums (Allen, 2006). SILCs in turn are groups formed by self-selected individuals and the members' contributions become a source of loans for improving food security, home repair, and the purchase of household and productive assets (Vanmeenen, 2010).

Some savings groups also provide informal insurance to group members. Savings group members contribute their savings to two funds: the money in the first fund is used to provide internal credits to members, while the second fund is a social fund aimed to improve members' welfare. The money in the social welfare fund typically covers costs and losses related to natural disasters and life-cycle events like marriages, funeral costs or the expenses of births (Andrew et al., 2018).

3.1 Paper A. Externalities of informal insurance: A theory-driven approach

Formal insurance and informal risk-coping strategies allow low-income households to face the costs of unexpected shocks caused by health emergencies or natural disasters. Informal risk-coping strategies include livelihood diversification (Ellis, 2000), migration (Skoufias, 2003), income diversification (Porter, 2012), changes in consumption patterns (Moser & Antezana, 2002), as well as selling assets, drawing on savings, or borrowing from the extended family in times of economic hardship (Strobl, 2017).

Another informal insurance strategy that offers to the poor an implicit coverage against idiosyncratic risks is the welfare fund of savings groups. The money in the social welfare fund typically covers the costs of losses related to life-cycle events like marriages, funerals, births or health-related issues (Dercon et al., 2006; Andrew et al., 2018).

Paper A evaluates the effect that the welfare fund has on loan allocation in savings groups. The theoretical mechanism proposed in **Paper A** suggests that behavioral changes arise from the higher affective and cognitive trust stimulated by the existence of a welfare fund in a group. Cognitive trust is related to the rational perception that, if a risk materializes, members do not have to face the financial consequences of paying for the cost of a loss event. Affective trust arises from the expectations of higher reciprocity among members during emergencies.

Viklund (2003) highlights that when group trust is higher, risk perception is lower. Positive affective trust particularly leads to lower risk perception and increased risk-taking according to Moreno et al. (2002). Brandstätter et al. (2002) add that in the presence of emotional trust, low probabilities of a loss event are overweighted and high probabilities are underweighted. Given this trust-based transmission channel, **Paper A** hypothesizes that the the informal insurance mechanism of the social welfare fund boosts loan allocation in savings groups.

Formally, let the difference between the estimators of theory-driven science and data-driven science be delineated on the basis of the Lemma 2 of **Box 1** in Section 2:

$$\hat{\Theta}_{\mathbb{H}} := \underset{\Theta \in \mathbb{R}}{\operatorname{argmin}}\, \ell\left(\mathbf{Y} - f_{\mathcal{S}}\left(\mathbf{X}, \hat{\Theta}\right)\right), \tag{3.1a}$$

$$\hat{\Theta}_{\mathbb{D}} := \underset{\Theta \in \mathbb{R}}{\operatorname{argmin}}\, \ell\left(\mathbf{Y}_i - f_{\mathcal{S},i}\left(\mathbf{X}_i, \hat{\Theta}_i\right), \mathbf{Y}_j - f_{\mathcal{S},j}\left(\mathbf{X}_j, \hat{\Theta}_j\right)\right). \tag{3.1b}$$

The estimators $\Theta_{\mathbb{H}}$ of the theory-driven approach—equation (3.1a)—are formulated based on a hypothesis $\mathbb{H}$ and the algorithmic optimization aims to minimize an in-sample loss function $\ell(\mathbf{Y} - f_{\mathcal{S}}(\mathbf{X}, \hat{\Theta}))$; this is, to maximize the in-sample fit. On the contrary, $\Theta_{\mathbb{D}}$—equation (3.1b)—aims to maximize predictive power by splitting the data (for $i \neq j$ partitions of $\mathbb{D}$). $\Theta_{\mathbb{D}}$ is the basis of machine-learning algorithms that minimize out-of-sample prediction or classification errors.

Paper A builds and tests a theory-driven estimator $\Theta_{\mathbb{H}}$ with the data on savings groups in the SAVIX. Conditional on other control covariates that can affect loan allocation—which are contained in an information set Ω_w—, **Paper A** hypothesizes that trust reduces risk perception and creates incentives for increasing loan allocation in savings groups:

$$\mathbb{H}_0 : \Theta_{\mathbb{H}} | \Omega_w > 0.$$

The hypothesis $\mathbb{H}_0$ of **Paper A** is tested with a correlational estimator $\theta_c \subset \Theta_{\mathbb{H}}$ and a quasi-experimental estimator $\theta_q \subset \Theta_{\mathbb{H}}$, using information on 147,580 savings groups operating worldwide. The estimator θ_c is constructed using panel regressions. Quasi-experimental methods—propensity score matching and augmented inverse probability weighting (AIPW)—are used to estimate the causal effect of the social welfare fund on loan allocation with θ_q.

In **Paper A**, the estimates $\hat{\theta}_c = 1.10$ and $\hat{\theta}_q = 1.94$ were found to be statistically significant at a level of less than 1% . The AIPW estimate $\hat{\theta}_q$ indicates that groups without a welfare fund allocate on average 9.85 USD loans per member, while groups with a welfare fund allocate 11.79 USD loans per member, i.e. 9.85 USD plus the point effect estimate of $\hat{\theta}_q = 1.94$. **Paper A** concludes that in groups with a social welfare fund members take higher amounts of loans. The findings support the theory that the social fund in savings groups not only implicitly covers idiosyncratic risks, but also increases trust among risk-sharing members and stimulates loan allocation.

Paper A contributes to the literature on informal insurance at the bottom of the wealth pyramid (BoP) by suggesting that informal risk-coping strategies can lead to investment decisions. Alderman and Paxson (1992) argue that insurance at the BoP provides a consumption risk protection against income fluctuations. The results of **Paper A** indicate that, besides serving as a protection against income fluctuations, the implicit insurance mechanism of the welfare fund in savings groups stimulates the provision of loans, and thus promotes sustainable investment strategies. Sustainable investment strategies, coupled with informal insurance, allow low-income households to make safer investments in income-generating opportunities that reduce poverty and lead to enhanced wealth creation.

Paper A also advances the knowledge about informal insurance by suggesting that the welfare fund of savings groups offers to the poor a trust-based risk-coping strategy with no moral hazard and no adverse selection. Townsend (1995) and Morduch (2006) note that insurance for the poor is limited by moral hazard, adverse selection, lack of price discrimination, and failure to achieve economies of scale[1]. In contrast to the formal (micro-) insurance contracts for the poor, the informal insurance of the welfare fund does not create adverse incentives to incur a loss because the money in the fund is used ex post by a common decision of the members and is not explicitly intended to cover a specific pre-defined loss event. Adverse selection, in turn, does not arise in the implicit risk-coping strategies of savings groups because the contribution to the social fund (the premium) is mandatory and in an amount orthogonal to the risk of the members.

3.2 Papers B and C. Profit generation and social programs in savings groups: A data-driven approach

The profit-generating capacity of savings groups can be captured with returns on savings (ROS). ROS quantifies how savings are transformed into loans and thus provides a direct measure of the collective benefits—the profits—generated in these associations[2]. ROS can

[1]Moral hazard arises because the very fact of having a formal insurance may raise the probability of experimenting a loss if the insured has fewer motivations to avoid the loss event. Adverse selection, in turn, is the result of riskiest individuals/households being the ones most eager to purchase insurance in the first place. When insurers cannot anticipate who is most risky, lack of price discrimination may discourage safer clients from taking a formal insurance.

[2]Returns are calculated by adding the values of the cash kept in the cashbox, the bank balance, the property of the group at the end of the cycle, and the loans outstanding. From the sum of these values are subtracted the value of savings, the property of the group at the start of the cycle, and the debts of the group. ROS are calculated by dividing the returns by half of the value of savings in a group.

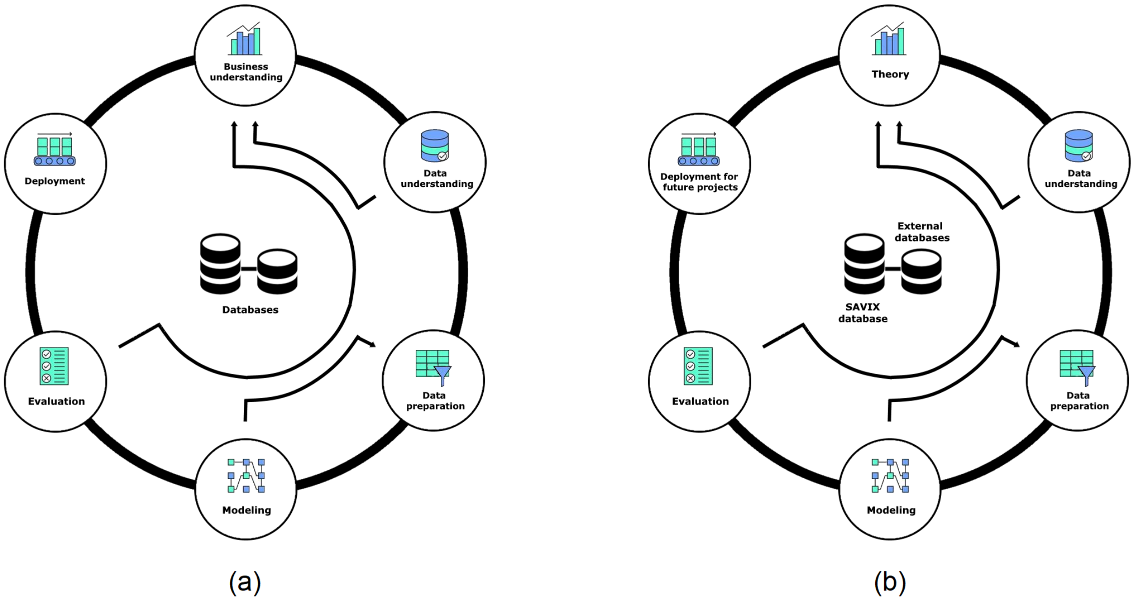


Figure 3.1: (a) Cross Industry Standard Process for Data Mining (CRISP-DM), and (b) CRISP-DM adapted for analyzing the SAVIX database with data-driven methods

be influenced by micro, meso, and macro factors. Micro factors are internal group-level characteristics. Macro factors reflect the macroeconomic environment in which a group operates. Meso factors are related to the facilitation process of development agencies that work with savings groups by providing them training and social programs.

Paper B applies Bayesian data-mining to identify which covariates—micro, meso and macro—are related to the profit-generating capacity of savings groups. Data-mining discovers patterns in large amounts of information, using data-driven statistical and mathematical techniques (Larose & Larose, 2014). Instead of being based on a single model that is vulnerable to specification errors, Bayesian data-mining applies Bayesian model-averaging to estimate thousands of models with different combinations of explanatory variables and then summarizes all the results to obtain the probability of each variable being related to profit generation. A similar approach is applied by *inter alia* Sala-i-Martin et al. (2004), Masanjala and Papageorgiou (2008) or Ley and Steel (2009). In **Paper B** the existing theories about savings groups are temporarily put it out of play—*außer spiel zu setzen*—as no hypothesis is formulated about a specific variable affecting profit generation. The potential explanatory variables of profits in savings groups are selected based only on previous empirical evidence and the available information in the SAVIX.

The data-driven process of **Paper B** is based on the Cross Industry Standard Process for Data Mining (CRISP-DM)—see Feyyad (1996). CRISP-DM starts with an understanding of the business and the data, after which data is cleaned to remove noise/outliers and the data is treated for dimensionality reduction and comparability (data preparation). In the fourth phase (discovery), modeling searches for patterns. In the fifth phase (evaluation), the mined patterns are interpreted, and in the last phase the model is deployed. The

The adjustment of the denominator to half of the savings is proposed in Rasmussen (2012, equation 5). Rasmussen (2012) argues that the division of the stock of savings by two brings ROS closer to the profile of profit generation in savings groups, because in these associations funds are available for loans not at the beginning of the cycle but rather at the end of a period of savings accumulation. Thus, on average, the funds available for loans at any moment of the cycle are half of the funds observed at the moment of data collection.

sequence of CRISP-DM is circular but not strict as the process moves moving back and forth between different phases (Figure 3.1a).

Paper B adapts the traditional CRISP-DM process to analyze the SAVIX. First, the SAVIX database is combined with external databases: the World Bank's World Development Indicators, and the World Risk Report published by the Alliance Development Works/Bündnis Entwicklung Hilft (BEH), the United Nations University Institute for Environment and Human Security (UNU-EHS). Second, theory/previous empirical evidence about savings groups is used instead of business understanding as guidance for data understanding and for evaluation of the results. Finally, after coherent results are obtained in the evaluation phase, a cleaned version of the SAVIX is deployed for future research projects (Figure 3.1b).

In **Paper B**, the principle of indifference—also called principle of insufficient reason—is applied in order to assign the epistemic prior probability of each variable explaining profit generation in the Bayesian models. This is, after bracketing theory (*Einklammerung*), equivalent states of knowledge are assumed for each covariate, and hence an equivalent probability is assigned to each explanatory variable *a priori* in the data-mining algorithm.

The data-driven results of **Paper B** show that micro group-level characteristics are not the most important variables for explaining profits. External factors—the macroeconomic environment and the facilitation model of development agencies—explain almost 80% of the profit-generating capacity of savings groups. The results about the importance of external factors are unexpected and challenge the pre-conceptions about the importance of internal group-level characteristics—like the gender composition of a group or the member's attendance to meetings—which have been the focus of the research on savings groups—see for example Gash and Odell (2013) or Entz et al. (2016).

At the macro level, savings groups operating in countries with a rural and more dispersed population have a higher probability of positive returns, which suggests that savings groups are better adapted to serve the rural poor. The most important meso variables for profit generation are the type of development agency working with a group, group status—graduated—and the provision of additional (plus) development services.

Graduated groups show higher returns compared to groups under the active supervision of NGOs, suggesting that the facilitation model of development agencies may not be providing enough flexibility to supervised groups. This result is in line with Boonyabancha (2001), Danquah et al. (2018), Jahns-Harms and Wilson (2018), and Gugerty et al. (2019), who recognize that the type of facilitation model implemented by a NGO has a dissimilar impact on profit generation.

The higher returns of groups working with development agency B—compared to the negative effect on returns observed in groups working with agency D—can be explained by the prioritization of social outputs over financial gains in the case of agency D, or by the innovations in the basic savings group model implemented by agency B[3]. These explanations—obtained through eliminative induction and abductive reasoning—are gestaltic, as they depend upon the unexpected patterns and the stronger relationships discovered through data-driven methods in **Paper B**.

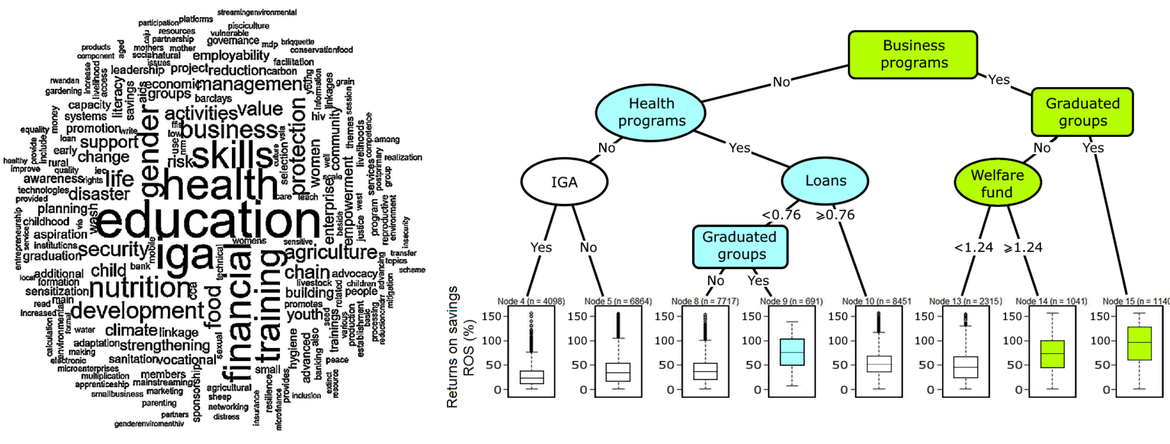
A deeper understanding of the role of development agencies is obtained in **Paper C**. **Paper C** applies data-driven methods based on enumerative inductive learning to

[3] Agencies D and B both follow a standardized model of savings groups that features 15 to 25 members, weekly meetings, an operational cycle of 9 to 12 months, and the integration of social development programs within a group. Agency B, however, offers more flexibility to the groups by implementing digital platforms that allow members to reduce the frequency of meetings from once a week to once every two weeks.

Table 3.1: Bayesian Data-Mining Results of **Paper B**

Potential explanatory variables of the profit generating capacity of savings groups	**Bayesian point estimates**	**Probability of driving profits**	**Importance for profit generation (%)**
Micro variables (group-level characteristics)			**21.06**
Members' attendance	0.0006	2.3273	0.3100
Dropout rate	0.0104	5.3539	1.1494
Number of loans outstanding	9.5990	55.7192	6.5517
Women members in a group	-0.0001	1.6563	0.0635
Group size	0.0055	3.5198	0.7553
Accumulated loans per member	7.5745	45.4549	5.3355
Welfare fund per member	0.4677	41.2770	4.5701
Fund utilization rate	-0.0064	9.4912	1.6251
Urban savings groups	0.0656	3.6412	0.6985
Meso variables (group facilitation by development agencies)			**52.00**
Provision of additional services	-0.5849	11.3580	1.8560
Additional social services	-6.4915	68.6892	7.6767
Additional financial services	1.0472	14.4235	2.1536
No facilitating agency	0.0117	2.7099	0.0866
Facilitating agency A	0.0515	2.6467	0.4308
Facilitating agency B	2.5021	26.5281	3.2839
Facilitating agency C	-1.3077	16.8697	2.4109
Facilitating agency D	-11.6309	96.1777	16.1800
Facilitating agency E	-0.4917	7.7627	1.4241
Group status: active	-3.3854	30.2865	3.9656
Group status: graduated	8.8459	72.2023	9.5944
Group status: supervised	-0.1522	3.7955	0.4794
Formation by field officer	-0.0148	2.2494	0.2860
Formation by village agent	-0.3563	3.1294	0.5738
Formation by group-paid agent	-0.0218	2.4676	0.3534
Formation by project-paid agent	-0.0664	2.7570	0.5488
Formation by unpaid agent	-0.0173	1.9471	0.1416
Formation by apprentice	-0.2029	2.7952	0.5143
Formation: spontaneous	-0.0093	2.2391	0.0445
Macro variables (macroeconomic environment)			**26.94**
Macroeconomic uncertainty	0.0032	2.2486	0.1160
Inflation rate	0.0006	2.4495	0.0633
Age-dependency ratio	0.0020	1.8517	0.1727
Inequality	-0.0233	3.4031	0.6604
Financial deepening	-0.0142	5.2619	1.0572
Literacy rate	-0.0134	3.8144	0.5963
GDP per capita	0.0006	6.1061	1.0036
Population density	-0.1987	96.0592	15.0254
Rural population	0.4897	46.5025	4.9910
Poverty (headcount rate)	-0.0139	3.0532	0.6720
GDP growth	0.1558	18.6345	2.5787

(*) In the data-mining models, control covariates were included to account for countries, back donors, between-cycles effects and within-cycle fixed effects. Back donors are institutions donating funds to development agencies, and include the African Development Bank, the Australian Agency for International Development (AusAID), the World Bank, the Bill & Melinda Gates Foundation, the Canadian International Development Agency (CIDA), the Inter-American Development Bank (IDB), the Master Card Foundation, and the United Nations Development Programme (UNDP).



(a) (b)

Figure 3.2: Inductive algorithms of data-driven science applied to the SAVIX. (a) The text-mining algorithm indicates that education, income-generating activities, and health programs are the most frequent programs provided by development agencies to savings groups. (b) Random forests show that business training is the most important social program for savings groups, particularly when group graduation is encouraged and there exist a social welfare fund in a group (green branch on the right side of the regression tree).

identify the most important social development programs for savings groups. **Paper C** randomly divides the SAVIX data into training and validation subsamples. Decision trees are estimated with the subset of training data and the combined result—the forest—is used to predict the target variable (returns on savings). **Paper C** takes an agnostic position by making a voluntary abstention (*enthaltung*) from development theory, neither postulating nor rejecting the importance of any particular social program.

The machine-learning results of **Paper C** identify two patterns in the SAVIX database: a text-mining algorithm finds that education, income-generating activities, and health programs are the most frequent programs provided by development agencies (Figure 3.2a); random forests identify business training as the most important social program for savings groups (Figure 3.2b). The green branch on the right side of the regression tree in Figure 3.2b shows that graduated savings groups that received business training (1149 groups) have a median ROS of 99%, with an interquartile range of 37% to 126%. Groups that received business training but are still supervised by a development agency have a median ROS of 75% if the group has a welfare fund equal to or greater than 1.24 USD.

The light-blue branch on the left side of the regression tree in Figure 3.2b shows that health development programs are the second-best predictor of returns in savings groups. Graduated groups that allocate loans and receive social programs related to health improvement have an average rate of returns on savings of 75%. If no health or business programs are offered to the groups, social programs aimed to promote income-generating activities (IGA) can also boost the financial performance of savings groups.

The data-driven findings of **Paper B** and **Paper C** can be used for theory formulation after the bracketing of theory is lifted, thus complementing the theory-driven approach in a circular fashion, as in the extension of the Eisenhardt (1989) model illustrated in Figure 2.1 of Chapter 2. As Chang et al. (2014) assert, one possible way to create theory more effectively is by iterating between data discovery and theory-building. In this case, theory-building—based on the data discoveries of **Papers B** and **C**—follows a process of *theoretical-model abduction*—which can be considered a pattern of creative abductive reasoning in the cubic model of Magnani (2011)—and is defined by Schurz (2008, 5, pp. 213) as the abductive task of finding theoretical conditions that describe the causes of the phenomenon in the theoretical language and that allow the mathematical derivation of the phenomenon from the theory.

Formally, a theoretical model of profit generation in savings groups—based on the data-driven findings of **Papers B** and **C**—can be framed under a 1-dimensional Itô process defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

$$S_t = S_0 + \int_0^t \mu(s, \omega)ds + \int_0^t \sigma(s, \omega)dB_s,$$

where $\{S_t\}_{t\in T}$ is a stochastic process that captures the dynamics of savings accumulation, $\omega \in \Omega$, $S_0 := S_{t=0} > 0$ are the savings of a group at the start of the cycle, and B_t is a Brownian motion—a Wiener process—related to the diffusion process $\sigma(s, \omega)$; see Oksendal (2013). In the Itô integral, $\mu(s, \omega)$ is a drift process that results from a convex conical combination of the external factors affecting the drift process (μ_e)—the macroeconomic environment and the facilitation mechanisms of development agencies—, and the internal factors affecting the drift process (μ_i), which are related to the internal dynamics of savings groups. These factors also affect the diffusion process through a convex conical combination of σ_e and σ_i. In shorter differential form, the Itô process can be written as:

$$dS_t = \left[\phi\mu_e + (1-\phi)\,\mu_i\right]dt + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]dB_t,$$

or:

$$\begin{cases} dS_t = \mu(s,t)dt + \sigma(s,t)dB_t \\ \mu(s,t) = \phi\mu_0 + (1-\phi)\mu_1 \\ \sigma(s,t) = \lambda\sigma_0 + (1-\lambda)\sigma_1 \end{cases},$$

in a state-space differential form, where $-\infty < \mu_e, \mu_i < \infty$, $\sigma_e > 0, \sigma_i > 0$. The parameters $\phi \in \mathbb{R}^{0,1}$, $\lambda \in \mathbb{R}^{0,1}$ are weights that measure the importance of external and internal factors for the drift (ϕ) and for the diffusion process (λ) of savings accumulation. The solution of the previous model can be found by applying Itô's lemma to the natural logarithm of S_t (Steele, 2012):

$$\begin{aligned} d(\ln S_t) &= \frac{\partial \ln S_t}{\partial t}\,dt + \frac{\partial \ln S_t}{\partial S}\,dS_t + \frac{1}{2}\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2 \frac{\partial^2 \ln S_t}{\partial S^2}\,dt \\ &= 0\,dt + \frac{1}{S_t}\,dS_t + \frac{1}{2}\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2 \left(-\frac{1}{S_t^2}\right)\,S_t^2\,dt \\ &= \frac{1}{S_t}\Big(\left[\phi\mu_e + (1-\phi)\,\mu_i\right]S_t\,dt + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]S_t\,dB_t\Big) - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2}\,dt \\ &= \left[\phi\mu_e + (1-\phi)\,\mu_i\right]\,dt + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]\,dB_t - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2}\,dt \\ &= \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2}\right)\,dt + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]\,dB_t. \end{aligned}$$

Integrating the last expression yields:

$$\int_0^t d(\ln S_t) = \int_0^t \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2} \right) \, ds + \int_0^t \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right] \, dB_s$$

$$\Rightarrow \ln S_t - \ln S_0 = \ln\left(\frac{S_t}{S_0}\right) = \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2} \right) \, t + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right] B_t.$$

The dynamics of savings can be written as

$$S_t = S_0 \; \exp \; \left\{ \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2} \right) t + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right] B_t \right\}.$$

Given that $B_t \sim \mathcal{N}(0,1)$, the dynamics of returns on savings is defined by $\frac{dS_t}{S_t}$:

$$\frac{dS_t}{S_t} = \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2} \right) \, dt + \left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right] \, d\left(\mathcal{N}(0,1)\sqrt{dt}\right).$$

Hence the returns $r_t := \frac{dS_t}{S_t}$ follow a Gaussian distribution with a mean equal to $\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e+(1-\lambda)\sigma_i\right]^2}{2}$ and a variance $\frac{\left[\lambda\sigma_e+(1-\lambda)\sigma_i\right]^2}{\sqrt{T}}$:

$$r_t \sim \mathcal{N}\left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{1}{2}\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2, \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{\sqrt{T}} \right).$$

The expected returns $\mathbb{E}(r_t)$ in savings groups are:

$$\mathbb{E}(r_t) = \left(\phi\mu_e + (1-\phi)\,\mu_i - \frac{\left[\lambda\sigma_e + (1-\lambda)\,\sigma_i\right]^2}{2} \right) \, dt.$$

Figure 3.3 shows simulations of the theoretical model of stochastic differential equations for savings groups, that is calibrated on the basis of the findings of **Papers B** and **C**. As suggested by the data-driven findings, external factors—the macroeconomic environment and the facilitation mechanisms of development agencies—are more important than internal group dynamics for profit generation, and hence in the model $\mu_e > \mu_i$ and $\sigma_e > \sigma_i$. Also, as external factors explain around 80% of ROS, then $\phi = \lambda = 0.8$. Figure 3.3c—which compares the model's simulations with the observed kernel density of ROS observed in the SAVIX database—suggests that the theoretical model—calibrated with the data-driven findings of **Papers B** and **C**—properly captures both the observed over-dispersion in the tails and the leptokurtosis in the distribution of ROS of the groups in the SAVIX database[4]. The data-driven evidence of **Paper B** and **Paper C** contributes this way to the theoretical formalization of the factors affecting the dynamics of savings groups.

Paper B and **Paper C** also contribute to the literature on savings groups by filling the

[4]Theoretical models, calibrated with data-driven evidence, improve the predictions of chaotic dynamics as those generated by the Kuramoto-Sivashinsky nonlinear partial differential equation. See for example Pathak et al. (2018), who hybridize the data-driven machine-learning approach with traditional knowledge-based models.

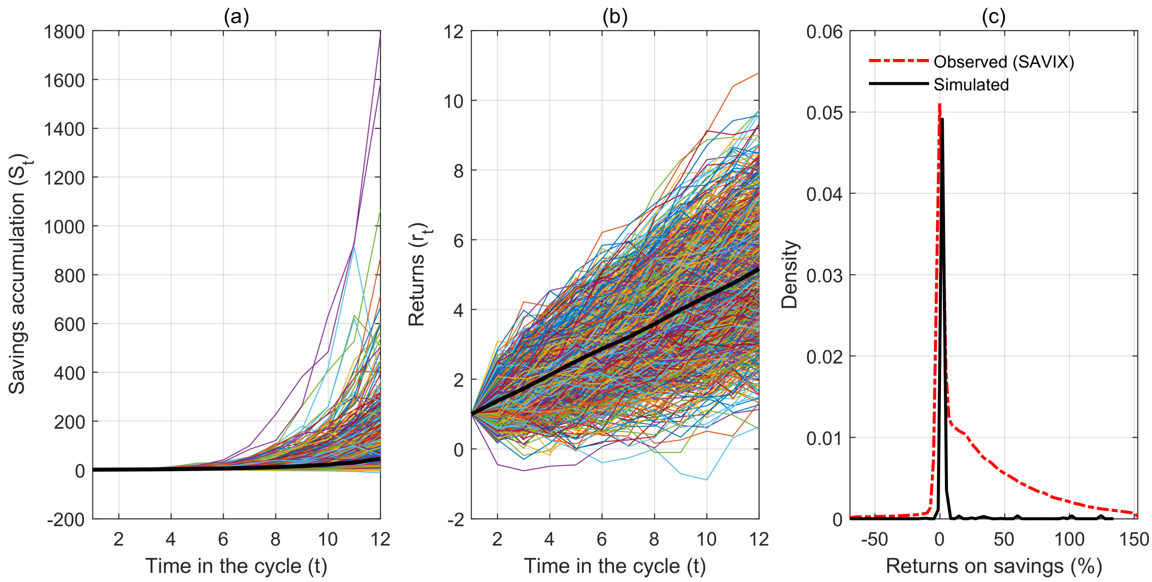


Figure 3.3: Simulation of the theoretical model of savings groups' dynamics based on the data-driven empirical findings of **Papers B** and **C**. (a) Simulation of the dynamics of savings accumulation of 1000 savings groups during a cycle of 12 months, (b) simulation of the dynamics of returns, and (c) density of returns on savings (ROS) in the observed SAVIX database (red line) and density of ROS simulated with the theoretical model (black line). In the stochastic differential equations of the model the constants are equal to $\mu_e = .55$, $\mu_i = .25$, $\sigma_e = .51$, $\sigma_i = .35$ and $\phi = \lambda = 0.8$.

research gap left by qualitative and context-specific studies that produced results that are not generalizable beyond the local context[5]. Armendáriz and Morduch (2010) argue that qualitative and context-specific studies are not a substitute for statistical evidence based on large samples, such as the evidence presented in **Papers B** and **C**. Instead of focusing on a small set of testable hypotheses—as previous studies framed in a theory-driven approach have done—**Papers B** and **C** use a data-driven approach to disentangle the differential effects that the micro, meso, and macro environments have on savings groups.

In practice, the results of **Papers B** and **C** provide valuable guidance to the efforts of back donors and development agencies working with savings groups. The findings of these papers indicate that these organizations should take into account that the macroeconomic environment, the facilitation mechanisms and the type of social program offered to savings groups can be more relevant than group-level characteristics for promoting financial inclusion and sustainable development at the bottom of the wealth pyramid.

3.3 Paper D. Simulating savings groups with artificial agents: Data-intensive science with complex algorithms

Computational simulations with complex algorithms are a new style of scientific reasoning that can be added to the list of scientific styles proposed by Hacking (2009) and Crombie and Shea (1995). Under this paradigm, data science not only analyzes *observational* data,

[5]Flynn and Sumberg (2018) for example use interviews to analyze the activities of 57 members of youth savings groups in Tanzania, Uganda, Zambia, and Ghana, while le Polain et al. (2018) use qualitative methods to analyze savings groups in the Congo.

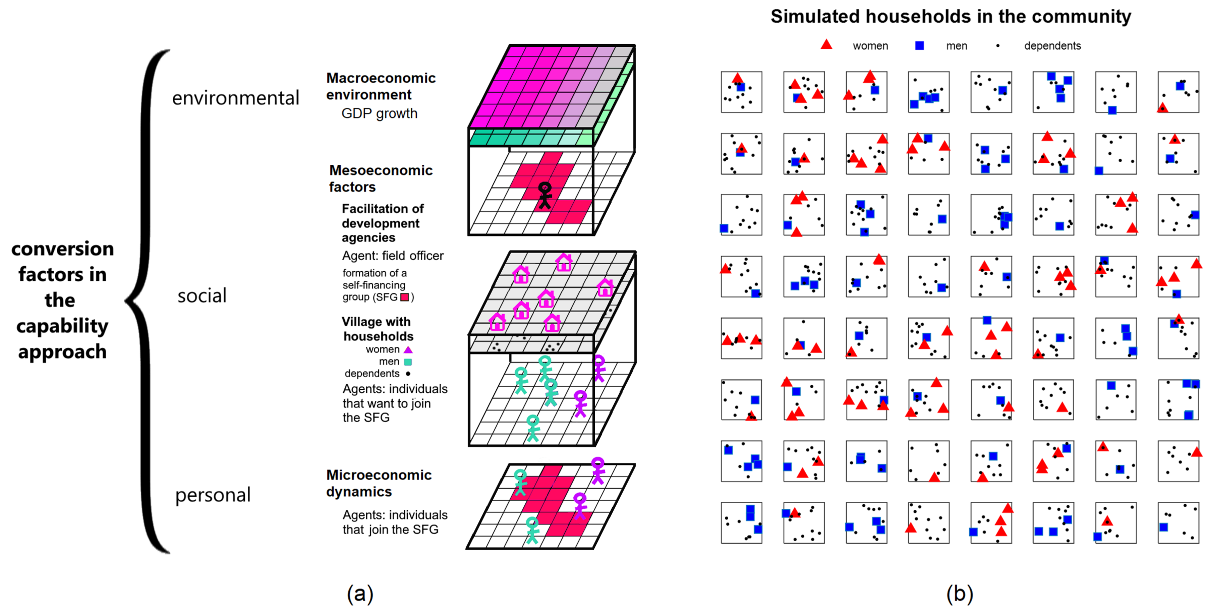


Figure 3.4: Virtual laboratory of savings groups. Figure (a) shows the multi-layered structure of the model in **Paper D**, which takes into account the macro, meso and micro factors affecting savings groups. This factors can be linked to the conversion factors of resources into functionings—beings and doings—in the Sen-Nussbaum capability approach. Figure (b) is an example of a village simulated with the complex algorithms of the model in **Paper D**.

but also generates and analyzes simulated data—this is, artificial data—based on theory. This new type of scientific reasoning was anticipated by Ruphy (2011, footnote 3).

In **Paper D**, complex algorithms are used to *create* large amounts of artificial data about savings groups. In this approach—based on agent-based modeling, see *inter alia* Lemos (2017), Hamill and Gilbert (2016), Rebaudo et al. (2011) or Sengupta and Bennett (2003)—, the dynamics of savings groups arise from the interaction between external factors—the macroeconomic environment and the facilitation mechanisms of development agencies—and the internal behavior of the individuals in the artificial community and the artificial savings group.

The model in **Paper D** is a sequence of four algorithms (Figure 3.4a). Algorithm 1 randomly creates an artificial population inhabiting households in a simulated village. In Algorithm 2, an agent hired by a development agency forms (artificial) savings groups. Algorithm 3 simulates the dynamics of the group. Finally, algorithm 4 simulates the financial performance of a business started by the members of savings groups. Taken together, the four algorithms create an artificial reality, this is, a microverse of agents that behave with artificial intelligence, as in Guterman et al. (2015)[6].

The multi-layered model of Figure 3.4 has a link with the conversion factors of the capability approach (Sen, 1976). In the capability approach, the real opportunity of an individual to do or achieve something depends on the conversion of resources into functionings—this is, states of human beings and activities. The conversion factors can be personal—internal microeconomic characteristics of a person, like age or gender—, social (mesoeconomic), and environmental (macroeconomic) (Robeyns, 2003). Just as in the capability approach, in the agent-based model of savings groups, people exercise agency,

[6]The oxymoron 'artificial reality' is not accidental: **Paper D** formulates a virtual laboratory that simulates artificial realities as a tool for experimenting with the behavior of savings groups.

which is the freedom and the ability to choose from the available options to pursue one's goals (Sen, 1993), which are in turn shaped by the specific socio-cultural context in which a person lives (Mchome et al., 2020).

The model of **Paper D** is as well inspired by the second order *simulacra* of Baudrillard (1994)—who motivated the simulated reality of Wachowski and Wachowski (1999). **Paper D** creates a virtual laboratory that emulates the behavior of members of savings groups in an artificial village (Figure 3.4b). The simulations of the virtual laboratory show that the businesses of the members of savings groups have higher profits due to the consolidation of social capital and the competitive advantage created through a dual process of homophily. This result complements the findings in **Paper C** about the importance of social programs related to business and entrepreneurship for savings groups.

In **Paper D**, the startup businesses of savings groups are more profitable and less risky compared to businesses financed with commercial loans, when social capital is properly consolidated. The consolidation of social capital is a consequence of the interaction among agents in the group. Social capital complements the debt capital in the fund available for loans, creating a competitive advantage that increases business profitability.

Higher quotas of savings in the group were found to boost profitability by raising the collective fund available for loans, but only up to a threshold, after which a bifurcation in returns appears[7]. In practice, the bifurcation implies that field officers—hired by development agencies for the task of managing a group—face a trade-off between two possible states when raising the savings quota of a savings group: while the bifurcation parameter is a potential source of profit, increasing the quota of savings exacerbates the risk of group failure.

The contributions of **Paper D** to previous research about savings groups is twofold. First, it provides quasi-teleological explanations about savings groups formation and performance, which are formalized in the complex algorithms of the model and create emerging patterns of savings groups' dynamics: community agents—the working-age population in the artificial village—decide to join a savings group in order to have access to social and financial benefits. Agents in charge of forming savings groups select group members in order to comply with a gender rule (a preference for women). The interaction between these dynamics creates emerging patterns not pre-defined in the model, as, for example, the number of members in a group or those members that fail to provide their quota of savings.

In practice, **Paper D** also makes a contribution by providing—to academics, donors and development agencies—a virtual laboratory of savings groups. As noted by Kort et al. (2003), virtual laboratories have the potential to become new important research tools, particularly if reactions of people to virtual environments are similar to those observed in real environments.

The virtual laboratory of **Paper D** can be used to perform counterfactual computational experiments of intervention programs in a community. Development agencies working with savings groups as a platform to provide sustainable development programs—like entrepreneurship, agriculture, adaptation to climate change, health and sanitation, or programs of literacy, education, and women's empowerment—can use and adapt the model in **Paper D** to perform artificial experiments. The impact of intervention programs and social policies can be evaluated ex ante through artificial experiments in the

[7]The bifurcation—typical in complexity dynamics (Gao et al., 2016)—is a branching process of the dynamical system in which the topological structure switches to different states due to a change in a bifurcation parameter (Crawford, 1991). In the model in **Paper D**, the bifurcation parameter is the quota of savings agreed among group members.

virtual laboratory. The simulation of interventions—before they are implemented—is a cost-effective way of reducing the risk of failure by taking into account the impact of unexpected (random) events and potential issues arising from policies implemented in the field.

Chapter 4

Conclusion

The availability of large-sample databases, increasing computational power, and the prominence of new statistical methods are changing scientific practice and are creating a momentum to consolidate the epistemology of data science (Leonelli, 2014). Using applications to a database of savings groups and simulations of artificial realities with complex algorithms, this dissertation contributes to this momentum by showing how data-driven methods complement the theory-driven approach.

The results of the applications in savings groups contribute to the literature on informal finance at the bottom of the wealth pyramid. Previous studies about financial instruments for the extreme poor have focused on the formal, supply-driven provision of microfinance arrangements. **Paper A to D** on the contrary provide new evidence about the internal dynamics of savings groups and the external factors that affect the performance of these informal financial associations. In practice, **Papers A to D** shed light on the best policies to promote sustainable development with informal finance: specifically, **Papers A to D** indicate that academics and practitioners working with savings groups have to prioritize the provision of business interventions and take into account the potential externalities caused by the social welfare fund, the macroeconomic environment of the country in which a group operates, and the facilitation model of development agencies.

Metatheoretically, the dissertation proposes a phenomenological framework for the epistemology of data-driven science. Based on a transitory abstention of theory, data-driven methods produce gestaltic and quasi-teleological explanations that can be used to formulate theory through inductive and abductive reasoning. The complementarity of data-driven and theory-driven science is formalized by extending the cyclical process of Eisenhardt and Graebner (2007).

Future studies can discuss the ontology, ethics, and teleology of data-driven methods. Investigating savings groups and the potential changes in their mechanics with theory-driven or data-driven methods—in virtual or real environments, with observed, experimental or artificial data—is a promising research avenue to promote sustainable financial inclusion and development in a cost-effective way.

Bibliography

Abbink, K., Irlenbusch, B., & Renner, E. (2006). Group size and social ties in microfinance institutions. *Economic Inquiry*, *44*(4), 614–628.

Alderman, H., & Paxson, C. H. (1992). *Do the poor insure?: A synthesis of the literature on risk and consumption in developing countries* (Vol. 164). World Bank Publications.

Allen, H. (2006). Village savings and loans associations — sustainable and cost-effective rural finance. *Small Enterprise Development*, *17*, 61–68.

Allen, H., & Panetta, D. (2010). Savings groups: What are they. *Washington DC: SEEP Network*, 2.

Alpaydin, E. (2016). *Machine learning: The new ai*. MIT Press.

Anderson, S., & Baland, J.-M. (2002). The economics of roscas and intrahousehold resource allocation. *The Quarterly Journal of Economics*, *117*(3), 963–995.

Andrew, D., George, J., & Sekei, L. H. (2018). Insights on the preferences and usage of financial services by savings groups in tanzania. *SatF Consortium*.

Anku-Tsede, O. (2013). Susu: A dynamic microfinance phenomenon in ghana. *Journal of Economics and Sustainable Development*, *4*(3), 146–155.

Ardener, S. (1964). The comparative study of rotating credit associations. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, *94*(2), 201–229.

Armendáriz, B., & Morduch, J. (2010). *The economics of microfinance*. MIT press.

Bacon, F. (1620). *Novum organum*. Clarendon press.

Bamford, G. (1993). Popper's explications of ad hoc ness: Circularity, empirical content, and scientific practice. *The British Journal for the Philosophy of Science*, *44*(2), 335–355.

Baudrillard, J. (1994). *Simulacra and simulation*. University of Michigan press.

Bauwens, L., Lubrano, M., & Richard, J.-F. (2000). *Bayesian inference in dynamic econometric models*. Oxford University Press.

Bell, R. M., & Koren, Y. (2007). Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, *9*(2), 75–79.

Bisrat, A., Kostas, K., & Feng, L. (2012). Are there financial benefits to join roscas? empirical evidence from equb in ethiopia. *Procedia Economics and Finance*, *1*, 229–238.

Blake, J. A., & Bult, C. J. (2006). Beyond the data deluge: Data integration and bio-ontologies. *Journal of biomedical informatics*, *39*(3), 314–320.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Psychology Press.

Boonyabancha, S. (2001). Savings and loans; drawing lessons from some experiences in asia. *Environment and Urbanization*, *13*(2), 9–21.

Bradshaw, G. F., Langley, P. W., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science*, *222*(4627), 971–975.

Brandstätter, E., Kühberger, A., & Schneider, F. (2002). A cognitive-emotional account of the shape of the probability weighting function. *Journal of Behavioral Decision Making*, *15*(2), 79–100.

Broad, C. (1968). The principles of demonstrative induction, In *Induction, probability, and causation*. Springer.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1–42.

Carnap, R. (1945). The two concepts of probability: The problem of probability. *Philosophy and phenomenological research*, *5*(4), 513–532.
Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, *63*, 67–80.
Chen, P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–65.
Crawford, J. D. (1991). Introduction to bifurcation theory. *Review of Modern Physics*, *63*, 991–1037.
Crombie, A., & Shea, W. (1995). Styles of scientific thinking in the european tradition. *Annals of Science*, *52*(6), 615–618.
Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 183–192.
Danquah, J. K., Analoui, F., & Koomson, Y. E. D. (2018). An evaluation of donor agencies' policies on participatory development: The case of ghana. *Development Policy Review*, *36*, O138–O158.
Deng, L., Yu, D. Et al. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387.
Dercon, S., De Weerdt, J., Bold, T., & Pankhurst, A. (2006). Group-based funeral insurance in ethiopia and tanzania. *World development*, *34*(4), 685–703.
Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.
Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766.
Drummond, J. J. (2007). *Historical dictionary of husserl's philosophy* (Vol. 81). Scarecrow Press.
Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review*, *14*(4), 532–550.
Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of management journal*, *50*(1), 25–32.
Ellis, F. (2000). The determinants of rural livelihood diversification in developing countries. *Journal of agricultural economics*, *51*(2), 289–302.
Elragal, A., & Klischewski, R. (2017). Theory-driven or process-driven prediction? epistemological challenges of big data analytics. *Journal of Big Data*, *4*(1), 19.
Empiricus, S. (2000). *Sextus empiricus: Outlines of scepticism*. Cambridge University Press.
Entz, M., Karsgaard, J., & Salomons, M. (2016). *An overview of savings and self-help groups, their contributions to improved food security, and how to improve their function: A review of the literature* (tech. rep.). Canadian Foodgrains Bank.
Evered, R. D. (1976). A typology of explicative models. *Technological Forecasting and Social Change*, *9*(3), 259–277.
Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. Humanities Press.
Feyyad, U. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE expert*, *11*(5), 20–25.
Flynn, J., & Sumberg, J. (2018). Are savings groups a livelihoods game changer for young people in africa? *Development in Practice*, *28*(1), 51–64.
Gao, J., Barzel, B., & Barabási, A.-L. (2016). Universal resilience patterns in complex networks. *Nature*, *530*(7590), 307.
Gash, M., & Odell, K. (2013). *The evidence-based story of savings groups: A synthesis of seven randomized control trials* (tech. rep.). SEEP Network.
Geertz, C. (1962). The rotating credit association: A middle rung in development. *Economic development and cultural change*, *10*(3), 241–263.
Ghosh, S., & Vinod, D. (2017). What constrains financial inclusion for women? evidence from indian micro data. *World Development*, *92*, 60–81.

Gill, J. (2015). *Bayesian methods: A social and behavioral sciences approach* (Vol. Third edition). CRC press.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science*, *17*(4), 319–321.

Good, I. J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, *30*(1/2), 39–73. http://www.jstor.org/stable/20115014

Gugerty, M. K., Biscaye, P., & Leigh Anderson, C. (2019). Delivering development? evidence on self-help groups as development intermediaries in south asia and africa. *Development Policy Review*, *37*(1), 129–151.

Guha, S., & Gupta, G. (2005). Microcredit for income generation: The role of rosca. *Economic and political weekly*, 1470–1473.

Guterman, D., Harmon, D., & Roiland, J. (2015). The ricks must be crazy [S(2), 6].

Hacking, I. (2009). *Scientific reason*. National Taiwan University Press Taipei.

Hamill, L., & Gilbert, N. (2016). *Agent-based modelling in economics* (1st). Wiley Publishing.

Handa, S., & Kirton, C. (1999). The economics of rotating savings and credit associations: Evidence from the jamaican partner. *Journal of development Economics*, *60*(1), 173–194.

Harper, W. L. (2011). *Isaac newton's scientific method: Turning data into evidence about gravity and cosmology*. Oxford University Press.

Hawthorne, J. (1993). Bayesian induction is eliminative induction. *Philosophical Topics*, *21*(1), 99–138. http://www.jstor.org/stable/43154143

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, *15*(2), 135–175.

Hendry, D., & Doornik, J. (2014). *Empirical model discovery and theory evaluation: Automatic selection methods in econometrics*. MIT Press.

Hey, T., Tansley, S., Tolle, K. M., Et al. (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Holloway, J., Mengersen, K., & Helmstedt, K. (2018). Spatial and machine learning methods of satellite imagery analysis for sustainable development goals. *16th Conference of the International Association of Official Statisticians (IAOS), 19-21 September, Paris, France*.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257.

Huzurbazar, V. (1955). On the certainty of an inductive inference, In *Proceedings of the cambridge philosophical society*.

Jahns-Harms, E., & Wilson, K. (2018). Let's look before we leap: Challenging our ideas on how savings groups work. *Enterprise development & microfinance*, *29*(3), 287–302.

Janowicz, K., Van Harmelen, F., Hendler, J. A., & Hitzler, P. (2015). Why the data train needs semantic rails. *AI Magazine*, *36*(1), 5.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Karlan, D., Savonitto, B., Thuysbaert, B., & Udry, C. (2017). Impact of savings groups on the lives of the poor. *Proceedings of the National Academy of Sciences*.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, *59*(7), 613–620.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481.

Kodratoff, Y. (2001). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. In G. Paliouras, V. Karkaletsis, & C. D. Spyropoulos (Eds.), *Machine learning and its applications: Advanced lectures* (pp. 1–21).

Kort, Y. A. d., Ijsselsteijn, W. A., Kooijman, J., & Schuurmans, Y. (2003). Virtual laboratories: Comparability of real and virtual environments for environmental psychology. *Presence: Teleoperators & Virtual Environments*, *12*(4), 360–373.

Kreif, N., & DiazOrdaz, K. (2019). Machine learning in policy evaluation: New tools for causal inference. *arXiv preprint arXiv:1903.00402*.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). The University of Chicago Press.

Lakatos, I., Worrall, J., & Currie, G. (1980). The methodology of scientific research programmes: Volume 1: Philosophical papers.
Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (Vol. 4). John Wiley & Sons.
le Polain, M., Sterck, O., & Nyssens, M. (2018). Interest rates in savings groups: Thrift or threat? *World Development*, *101*, 162–172.
Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.
Lemos, C. M. (2017). *Agent-based modeling of social conflict: From mechanisms to complex behavior* (1st). Springer Publishing Company, Incorporated.
Leonelli, S. (2014). What difference does quantity make? on the epistemology of big data in biology. *Big data & society*, *1*(1), 2053951714534395.
Lewis, D. (1973). Causation. *The journal of philosophy*, *70*(17), 556–567.
Ley, E., & Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 651–674.
Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width, In *Advances in neural information processing systems*.
Luft, S. (2004). Husserl's theory of the phenomenological reduction: Between life-world and cartesianism. *Research in Phenomenology*, *34*, 198–234.
Mackey, L. W., Syrgkanis, V., & Zadik, I. (2017). Orthogonal machine learning: Power and limitations. *CoRR*, *abs/1711.00342*arXiv. http://arxiv.org/abs/1711.00342
Magnani, L. (2011). *Abduction, reason and science: Processes of discovery and explanation*. Springer Science & Business Media.
Manin, Y. I. (2013). Kolmogorov complexity as a hidden factor of scientific discourse: From newton's law to data mining. *arXiv preprint arXiv:1301.0081*.
Markus, M. L. (2014). Maybe not the king, but an invaluable subordinate: A commentary on avison and malaurent's advocacy of 'theory light' is research. *Journal of Information Technology*, *29*(4), 341–345.
Masanjala, W. H., & Papageorgiou, C. (2008). Rough and lonely road to prosperity: A reexamination of the sources of growth in africa using bayesian model averaging. *Journal of applied Econometrics*, *23*(5), 671–682.
Massie, P. (2013). Philosophy and ataraxia in sextus empiricus.
Mazzocchi, F. (2015). Could big data be the end of theory in science?: A few remarks on the epistemology of data-driven science. *EMBO reports*, *16*(10), 1250–1255.
Mchome, Z., Yousefzadeh, S., Bailey, A., & Haisma, H. (2020). when i breastfeed, it feels as if my soul leaves the body": Maternal capabilities for healthy child growth in rural southeastern tanzania. *International Journal of Environmental Research and Public Health*, *17*(17), 6215.
McKeever, K. (2009). A short history of tontines. *Fordham Corporate and Financial Law Review*, *15*, 491–521.
Moran, D., & Cohen, J. (2012). *The husserl dictionary*. Bloomsbury Publishing.
Morduch, J. (2006). Microinsurance: The next revolution. *Understanding poverty*, 337–356.
Moreno, K., Kida, T., & Smith, J. F. (2002). The impact of affective reactions on risky decision making in accounting contexts. *Journal of Accounting Research*, *40*(5), 1331–1349.
Moser, C., & Antezana, O. (2002). Social protection in bolivia: An assessment in terms of the world bank's social protection framework and the prsp. *Development Policy Review*, *20*(5), 637–656.
Niemeijer, D. (2002). Developing indicators for environmental policy: Data-driven and theory-driven approaches examined by example. *Environmental Science & Policy*, *5*(2), 91–103.
Oksendal, B. (2013). *Stochastic differential equations: An introduction with applications*. Springer Science & Business Media.
Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2018). Orthogonal random forest for causal inference. *arXiv preprint arXiv:1806.03467*.

Parish, E. J., & Duraisamy, K. (2016). A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, *305*, 758–774.

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(4), 041101.

Peirce, C. S. (1960). *Collected papers of charles sanders peirce*. Harvard University Press.

Piekkari, R., & Welch, C. (2011). *Rethinking the case study in international business and management research*. Edward Elgar Publishing.

Piekkari, R., & Welch, C. (2018). The case study in management research: Beyond the positivist legacy of eisenhardt and yin. *The Sage handbook of qualitative business and management research methods*, 345–359.

Popper, K. R. (1963). Science as falsification. *Conjectures and refutations*, *1*, 33–39.

Porter, C. (2012). Shocks, consumption and income diversification in rural ethiopia. *Journal of Development Studies*, *48*(9), 1209–1222.

Rasmussen, O. D. (2012). Small groups, large profits: Calculating interest rates in community-managed microfinance. *Enterprise Development and Microfinance*, *23*.

Rebaudo, F., Crespo-Pérez, V., Silvain, J.-F., & Dangles, O. (2011). Agent-based modeling of human-induced spread of invasive species in agricultural landscapes: Insights from the potato moth in ecuador. *Journal of Artificial Societies and Social Simulation*, *14*(3), 7.

Robeyns, I. (2003). Sen's capability approach and gender inequality: Selecting relevant capabilities. *Feminist economics*, *9*(2-3), 61–92.

Ruphy, S. (2011). From hacking's plurality of styles of scientific reasoning to "foliated" pluralism: A philosophically robust form of ontologico-methodological pluralism. *Philosophy of science*, *78*(5), 1212–1222.

Sala-i-Martin, X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American economic review*, 813–835.

Schurz, G. (2008). Patterns of abduction. *Synthese*, *164*(2), 201–234.

Seel, G. (2018). *Where next after vsla? a world renew research paper exploring the value of linking savings groups with financial services providers* (tech. rep.). World Renew.

Semenova, V. (2018). *Essays in econometrics and machine learning* (Doctoral dissertation). Massachusetts Institute of Technology.

Sen, A. (1976). Famines as failures of exchange entitlements. *Economic and Political Weekly*, 1273–1280.

Sen, A. (1993). Capability and well-being73. *The quality of life*, *30*, 270–293.

Sengupta, R. R., & Bennett, D. A. (2003). Agent-based modelling environment for spatial decision support. *International Journal of Geographical Information Science*, *17*(2), 157–180.

Skoufias, E. (2003). Economic crises and natural disasters: Coping strategies and policy implications. *World development*, *31*(7), 1087–1102.

Solomonoff, R. J. (1964). A formal theory of inductive inference. part ii. *Information and control*, *7*(2), 224–254.

Steele, J. M. (2012). *Stochastic calculus and financial applications* (Vol. 45). Springer Science & Business Media.

Strobl, R. (2017). Does health insurance reduce child labour and education gaps? evidence from rwanda. *The Journal of Development Studies*, *53*(9), 1376–1395.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis: Theorizing qualitative research*. University of Chicago Press.

Townsend, R. M. (1995). Consumption insurance: An evaluation of risk-bearing systems in low-income economies. *Journal of Economic perspectives*, *9*(3), 83–102.

Van Maanen, J., Sørensen, J. B., & Mitchell, T. R. (2007). The interplay between theory and method. *Academy of management review*, *32*(4), 1145–1154.

Vanmeenen, G. (2010). *Savings and internal lending communities: Voices from africa*. Catholic Relief Services.

Viklund, M. J. (2003). Trust and risk perception in western europe: A cross-national study. *Risk Analysis: An International Journal*, *23*(4), 727–738.

Wachowski, L., & Wachowski, L. (1999). The matrix [W. B. and Roadshow E.].

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zabell, S. L. (2011). Carnap and the logic of inductive inference, In *Handbook of the history of logic*. Elsevier.

Appended Papers

A Informal insurance and loan allocation in savings groups: the role of the welfare fund

B What drives profit generation in savings groups? Bayesian data-mining discoveries

C Which social program supports sustainable grassroot finance? Machine-learning evidence

Published in: International Journal of Sustainable Development & World Ecology.
Volume 27, 2020 - Issue 5, pages 389-395.

Abstract

Resources for sustainable development are used efficiently when social programs help to promote simultaneously the financial sustainability of grass-root groups that provide financial access to millions of low-income households around the world. This study applies machine-learning to a worldwide database of grass-root groups in order to identify which social programs are good predictors of financial returns in the groups. The results indicate that education, income-generating activities and health programs are the most frequent programs provided by development agencies. Business training is not among the most frequent interventions applied in grass-root groups, but it is in fact the most important social program to encourage financial sustainability in grass-root groups, particularly after a development agency stops working with a group and leaves the community. Theoretical and practical implications of the findings are discussed.

C.1 Introduction

Financial access to low-income households in developing countries tends to be provided by microfinance institutions. In contrast to this institutional approach, in grass-root finance individuals living in impoverished communities create a group and start to accumulate their savings into a fund, which is later used to provide small loans to themselves.

Grass-root groups receive different names, like inter alia savings groups (Allen & Panetta, 2010), self-help groups (Venkatraja, 2019), rotating savings and credit associations or accumulating savings and credit associations (Bouman, 1995). Greaney et al. (2016) and Burlando and Canidio (2017) estimate that over 100 million persons in 10.5 million households participate in grass-root financial groups worldwide.

Studying the dynamics of informal grass-root finance is extremely important due to the recent evidence that expanding the access to formal savings and loans will not be enough to broaden financial access to the poor (Dupas et al., 2018). Grass-root finance arises at the bottom of the pyramid (BoP) and reaches the poorest population in developing countries, who do not have access to formal financial services and rely exclusively on grass-root finance to meet their needs (Burlando & Canidio, 2017).

International donors and development agencies have recognized the relevance of grass-root finance for poverty reduction. These organizations work with grass-root groups as a platform to provide communities with sustainable development programs. The agencies help a community to organize a financial group, and then provide a development service to the group, like entrepreneurship, agriculture, adaptation to climate change, health and sanitation, or programs of literacy, education and women empowerment.

Examples of grass-root financial associations supported by development agencies are the Village Savings and Loan Association (VSLA) promoted by CARE International or the Savings for Changes (SfC) model supported by Oxfam and Freedom from Hunger (Le Polain et al., 2018). During the implementation of social and financial programs, development agencies work closely with donors like the Inter-American Development Bank or the Barclays Corporation (Flynn, and Sumberg, 2017).

Given the wide variety of sustainable development programs that agencies and donors can provide to grass-root groups, the question arises as to which program helps to promote both the social and financial sustainability of a group.

A group is financially sustainable when generates returns during the process of savings accumulation and loan provision. Groups that generate returns will have incentives to keep operating over time and thus will continue providing financial access to bottom-of-the-pyramid (BoP) individuals. Groups that are financially sustainable may also opt to maintain their social programs alive, even after the development organization leaves the community.

This study uses machine-learning methods to identify which social development programs are the best predictors of financial returns generated by grass-root associations. Text-mining and random forests are applied to the SAVIX, a database with information of more than 250000 grass-root financial associations worldwide.

The results indicate that education, income-generating activities and health are the most frequent development programs provided to grass-root groups. Training to create small retail businesses and health interventions are provided less frequently, but interestingly these interventions are in fact the ones that boosts the profit-generating capacity of grass-root groups, particularly business training in graduated groups that are no longer supervised by a development agency.

In the practice, the results imply that donors and development agencies that look to achieve social targets but also want to support financial sustainability can prioritize the provision of health programs and business training to grass-root groups. In combination with social interventions, encouraging groups to 'graduate' and become an autonomous and unsupervised association further enhances financial sustainability.

Theoretically, the findings contribute to the business literature by suggesting that grass-root business stimulated by development interventions can be thought as a new paradigm of sustainable business, complementary to social enterprises and corporate social responsibility.

Next section describes the methods and data used in the study. Section 3 details the results and Section 4 concludes by discussing practical implications and theoretical contributions, as well as the links of the findings with the Sustainable Development Goals (SDGs).

C.2 Methods and data

Machine learning is applied to the SAVIX database in order to identify which sustainable development programs are the best predictors of the profit-generating capacity of grass-root financial organizations across the world.

The Savings Groups Information Exchange (SAVIX) database has information of 250000 grass-root financial groups in 52 countries around the world (Figure 4.5). The group-level data of the SAVIX is collected in the field through an online system, the Savings Groups Management Information System (MIS), which is supported by the Bill & Melinda Gates Foundation, CARE, Catholic Relief Services, Oxfam America and Plan International.

Machine-learning are supervised, semi-supervised or unsupervised algorithms that allow to make predictions and create data-driven knowledge from a database. Applications of machine-learning in sustainable development are promoted by the United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014). Recent applications of machine-learning include the identification of harmful environmental impacts caused by unsustainable business (Can & Alatas, 2017) and the application of machine-learning for monitoring the SDG indicators (Holloway et al., 2018).

In this study, two common machine learning methods are applied to the SAVIX: text-mining and random forests.

Text-mining is an unsupervised process that seeks to extract useful information and identify patterns in textual data (Feldman & Sanger, 2007). The text-mining implementation in this study transforms the unstructured information of social interventions in the SAVIX into a corpus, i.e. a collection of writing data about sustainable development programs. The corpus is processed and summarized into a matrix of tokens, which is analyzed to find word frequencies and patterns. The text-mining results are displayed with word-clouds, where the word size denotes the frequency of a word in the corpus—see Weiss et al. (2015), Vijayarani et al. (2015) or Zhou et al. (2016).

Random forest are supervised algorithms that fit decision-trees to random subsets of training data and use the combined result—the forest—for prediction (Breiman, 2001). Decision-trees split the dataset into smaller subsets, with the aim of increasing the predictive power of the model for the target variable (Genuer et al., 2017). In this study, the individual predictions from the trees estimated with the SAVIX are combined into a final prediction of financial returns obtained by the groups. The importance of each variable for returns is calculated with node impurity—a measure of the splits that have a high inter-node variance and a small intra-node variance— and with the increase in the mean squared error of predictions (see Gregorutti et al., 2017).

A final decision-tree is estimated with the full sample (train and test) in order to find out which development programs are the best predictors of the profit-generating capacity of grass-root financial associations.

C.3 Results

The SAVIX data shows that grass-roots finance is provided to the poorest population living in rural communities and urban slums in developing countries. In the database, 65% of the groups operate in rural regions, 33% in urban slums and only 2% in urban regions. The average amount of savings per member in the grass-root groups is 17.5 USD, with a median of 9.5 USD. The value of loans provided to the members is on average only 12 USD, with a median of 6 USD (Table 4.7). The low values of savings and loans show that grass-root finance provides financial services to the extreme poor in the BoP.

Despite the low values of savings and loans in grass-roots associations, these groups have on average returns on savings (ROS) equal to 45%, with a median of 36% (Figure 4.6). Allen and Panetta (2010) explain that the high financial returns of grass-roots associations is the consequence of groups charging monthly interest rates ranging from 5 to 10 percent. Moreover, Guha and Gupta (2005) note that because members must repay the loan and pay interests, as well as keep contributing with their savings, a surplus that boosts returns arises naturally in grass-root associations.

In the SAVIX, 57863 grass-roots groups have records of having received a social intervention from a development agency. The text-mining of the corpus of development interventions indicates that the most frequent programs provided to grass-root groups are related to education, income-generating activities and health (Table 4.8, Figure 4.7). Business training is not among the most frequent interventions in grass-root financial associations.

One-hot encoding is used to translate the corpus to a binary matrix of development programs. The matrix is included in the random-forest model to estimate which social interventions predict higher returns. In order to control for other variables that could affect profit generation, group-level characteristics and macro-economic variables are added as controls in the random forests[13].

The number of variables that are randomly selected for splitting at each node of the trees are selected with cross-validation[14]. In the cross-validation, the database is randomly divided in a validation and a train set, with 65% of the data in the train set. The minimum value of both the root mean squared error (RMSE) and the mean absolute prediction error (MAPE) is obtained with 5 splitting variables (Table 3). Five splitting variables also maximize the correlation of predictions in the test set against the estimations of ROS in the train set (Table 4.9 and Figure 4.6)[15].

[13]The number of loans per member is included as a control covariate of operating efficiency and the status of the group —supervised against graduated— is used to account for the stage of agency monitoring (Ledgerwood et al., 2013). The welfare fund of a group was included as a control because the social fund can act as a collateral mechanism to cope with risks. Population density and the percentage of rural population of a country are included as macroeconomic controls because in less populated rural areas transport costs limit the possibilities of members attending meetings, thus increasing the chances of members not contributing with their savings and/or not paying their debts to the group (Christensen, 1993).

[14]Nicodemus & Shugart (2007) and Strobl et al. (2008) highlight that the ability of random forests to detect influential predictor variables depends on the number of selected splitting variables.

[15]Out-of-bag (OOB) errors are also used to validate the random forest model. The OOB error is the average error calculated using bootstrapped predictions from the trees with a specific number of splitting variables randomly permuted during the estimation of the forest (Gregorutti et al., 2017). An OOB error of 601.56 is obtained with 4 splitting variables, 594.91 with 5 splitting variables and 597.26 with 6 splitting variables.

Figure 4.8 illustrates the results about the importance of each variable for profit generation obtained with a random forest that has 5 splitting variables in each node.

The highest increase in MSE and node impurity is obtained for the variable loans per member, indicating that loans are the main predictor of returns in grass-root financial associations. This result is expected, because loan allocation is the main channel of profit generation in grass-root associations. The fact that loans are identified as the best predictor of profit generation supports the ability of the machine-learning algorithm to truly detect predictors of financial returns.

The welfare fund is the second most important factor that predicts higher returns in grass-root financial associations (Figure 4.8). While the purpose of the welfare fund is to offer grants or interest-free loans to cover emergencies and life-cycle events, Maliti (2017) found that groups in Tanzania appeal to the welfare fund for loan repayment, a deviation from the original purpose of the welfare fund that can increase returns.

In relation to development interventions, agricultural programs and training in WASH (water, sanitation and hygiene) have the lowest effect on financial returns (Figure 4.8), while in contrast development programs for the creation of small retail businesses have the highest predictive power for the generation of returns in grass-root financial organizations.

Similar results are obtained with the final decision-tree estimated with the full dataset (Figure 4.9). Graduated grass-root groups that receive business training (1149 groups) have a median of ROS equal to 99%, with an interquartile range of 37% to 126% (Node 15, Figure 5). Groups that receive business training but are still supervised by a development agency have a median of ROS equal to 75%, if the group has a welfare fund equal or higher than 1.24 USD (Node 14, Figure 4.9). Groups that received business training, but are still under supervision and have a smaller amount of money in the welfare fund (less than 1.24 USD per member) have a median of returns on savings equal to 48% (Node 13, Figure 4.9).

Health development programs are the second-best predictor of returns in grass-root financial associations, when the groups manage to provide loans to their members. Graduated groups that have received training in health and at the same time provide loans have an average rate of returns equal to 75% (Node 9, Figure 4.9).

If no health or business programs are offered to the groups, social programs aimed to promote income generating activities can also boost the financial sustainability of the groups (Figure 4.9, Nodes 4 and 5), but on a lower magnitude compared to groups that received business training and to those that are no longer under the active supervision of a development agency.

C.4 Discussion

Higher returns in grass-root financial associations help to ensure sustainable finance and the continuity of business and social development interventions. Groups with higher profits will have more incentives to keep operating over time, and thus will keep providing informal financial services to the poor. Likewise, financially sustainable groups will keep supplying social development programs to vulnerable populations for longer periods of time.

The machine-learning results show that the most frequent interventions of development agencies are related to education, income-generating activities and health. Business training is not the most frequent development program offered to grass-root associations, but it is in fact the most important predictor of financial returns, particularly for graduated groups already trained by development organizations. In the practice, the findings imply that donors and managers of development agencies interested in both social and financial sustainability at the BoP can make an efficient use of scarce resources by implementing interventions of health and business training, because these programs will simultaneously promote both financial and social welfare of low-income individuals participating in grass-root associations.

The findings also indicate that the staff hired by development organizations have to provide enough training to allow groups to achieve graduation. The training staff has to encourage leadership and trust within a group, as well as proactive membership and the ability and motivation to follow group rules (Delany and Storchi, 2012).

The findings of the study also contribute to theory-generation by suggesting that grass-root business stimulated by development interventions can be thought as a new paradigm of sustainable businesses at the BoP, complementary to the traditional paradigms of social enterprises (Hossain et al., 2017) and corporate social responsibility (Kolk, 2016).

In the case of corporate social responsibility, Hoque et al. (2018) suggest that —both in developed and developing countries— this paradigm tends to be a voluntary philanthropic fashion not necessarily focused on improving social well-being but rather aimed to build public image and enhance business profit. The businesses stimulated by grass-root financial organizations are on the contrary created and owned by the members, and thus are primarily interested in improving the lives of the participants.

Social enterprises on the other hand can be imposed on the poor by governments and developing agencies. In contrast to this centralized and authoritarian business model, grass-root businesses spawned by grass-root financial associations answer to the necessities, possibilities and expectations of the members.

Social enterprises also tend to prioritize social value over economic value (Seelos & Mair, 2005), while in contrast the small businesses promoted by grass-roots groups focus on both, social and economic value.

Business creation through grass-root finance is a community model that helps households to smooth consumption and also aids individuals to self-finance their productive investments in human and business capital (Karlan and Zinman, 2014). In the businesses created through sustainable grass-root financial organizations, individuals raise capital from their community and invest their money locally. Since grass-root businesses are locally owned and locally financed, wealth-creation remains in the community.

Finally, this study has also implications for the efficient allocation of resources of governments and development agencies interested in achieving the SDGs. The development programs promoted through grass-root financial organizations are linked to the SDGs 1 (no poverty), 2 (zero hunger), 3 (health), 4 (education), 5 (gender equality), 6 (water and sanitation), 8 (decent work and economic growth) and 13 (climate action). The results show that grass-root groups that receive business training can achieve simultaneously multiple targets of the SDGs 8, 3 and 1.

Grass-root financial groups can also be a platform to implement other development interventions, as for example programs to reduce psychological aggression and/or physical punishment to children (SDG 16) or projects to provide sources of renewable energy (SDG 7)[16]. Future studies can explore the simultaneous social and financial impact of these interventions, in order to improve the efficient use of resources for development and create insights about additional opportunities to make a stronger impact on the well-being of people around the world.

Acknowledgements

The author would like to thank Roy Mersland, Bert D'Espallier, Linda Nakato and Annikka Berridge. This research was supported by funding from the FAHU Foundation.

References

Allen, H., and Panetta, D. (2010). Savings groups: What are they? Technical Report, SEEP Network.

Bouman, F. (1995). Rotating and accumulating savings and credit associations: A development perspective. World Development, 23.

[16]ARED (African Renewable Energy Distributor) for example, is a micro-franchising business at the BoP that offers solar-powered mobile kiosks in Africa (Gabriel, and Kirkwood, 2016). This business model can be combined with grass-root finance to achieve simultaneously multiple goals of the SDGs 7, 8 and 1.

Breiman, L. (2001). Random forests. Machine learning, 45, 5–32.

Burlando, A. and Canidio, A. (2017). Does group inclusion hurt financial inclusion? Evidence from ultra-poor members of Ugandan savings groups. Journal of Development Economics, 128, 24–48.

Christensen, G. (1993). The limits to informal financial intermediation. World Development, 21, 721–731.

Delany, Aislinn and Silvia Storchi (2012). SaveAct Savings and Credit Groups and Small Enterprise Development. In: FinMark Trust.

Dupas, P., Karlan, D., Robinson, J., & Ubfal, D. (2018). Banking the unbanked? evidence from three countries. American Economic Journal: Applied Economics, 10, 257–97. Feldman, R. and Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Flynn, J., Sumberg, J. (2017). Youth savings groups in Africa: they're a family affair. Enterprise Development and Microfinance, 28(3), 147-161.

Gabriel, C. A., & Kirkwood, J. (2016). Business models for model businesses: Lessons from renewable energy entrepreneurs in developing countries. Energy Policy, 95, 336-349.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C. and Villa-Vialaneix, N. (2017). Random forests for big data. Big Data Research, 9, 28–46.

Greaney, Brian P, Joseph P Kaboski, and Eva Van Leemput (2016). Can self-help groups really be self-help? The Review of Economic Studies 83.4, pp. 1614–1644.

Gregorutti, B., Michel, B. and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27, 659–678.

Guha, S., and Gupta, G. (2005). Microcredit for income generation: the role of RoSCA. Economic and political weekly, 1470-1473.

Hoque, N., Rahman, A. R. A., Molla, R. I., Noman, A. H. M. and Bhuiyan, M. Z. H. (2018). Is corporate social responsibility pursuing pristine business goals for sustainable development? Corporate Social Responsibility and Environmental Management, 25, 1130–1142.

Holloway, J., Mengersen, K., Helmstedt, K. (2018). Spatial and machine learning methods of satellite imagery analysis for sustainable development goals. 16th Conference of the International Association of Official Statisticians (IAOS), 19-21 September, Paris, France.

Hossain, S., Saleh, M. A., & Drennan, J. (2017). A critical appraisal of the social entrepreneurship paradigm in an international setting: a proposed conceptual framework. International Entrepreneurship and Management Journal, 13, 347–368.

Karlan, Ratan, L., & Zinman, J. (2014). Savings by and for the poor: A research review and agenda. Review of Income and Wealth, 60, 36–78.

Kolk, A. (2016). The social responsibility of international business: From ethics and the environment to CSR and sustainable development. Journal of World Business, 51, 23–34.

Ledgerwood, J., J. Earne, and C. Nelson (2013). The New Microfinance Handbook: A Financial Market System Perspective. The World Bank.

Le Polain, M., Sterck, O., & Nyssens, M. (2018). Interest rates in savings groups: Thrift or threat? World Development, 101, 162 − 172.

Maliti, E. (2017). Deviation of community savings groups from their apparent methodology: lessons from the field. International Journal of Social Economics, 44, 326–336.

Seelos, C., & Mair, J. (2005). Social entrepreneurship: Creating new business models to serve the poor. Business horizons, 48, 241–246.

United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014). A World that Counts Mobilizing the Data Revolution for Sustainable Development. Technical Report, Data Revolution Group. Venkatraja, B. (2019). Are SHGs catalysts of inclusive sustainable rural development? Impact assessment of SKDRDP interventions in India. International Journal of Sustainable Development and World Ecology, 26(4), 302-312.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5, 7–16.

Table 4.7: Descriptive statistics of the variables in the SAVIX database

	Minimum	Median	Mean	Maximum
Members attendance (%)	39.29	96.00	91.24	100
Dropout rate (%)	0	0	1.79	45
Number of loans outstanding per member	0	0.36	0.38	1
Women members in the group (%)	0	86.96	78.99	100
Group size (number of members)	5	21	21	34
Accumulated loans per member	0	0.30	0.75	133.33
Welfare fund per member (USD)	0	0.47	1.10	12.59
Fund utilization rate	0	40.44	41.25	100
Savings per member (USD)	0	9.52	17.46	139.21
Loan-value per member (USD)	0	5.70	12.06	116.22
Returns on savings (%)	1	35.71	45.33	156.12

Table 4.8: Text-mining results for the 12 more frequent words in the records of development services offered to grass-root financial organizations in the SAVIX. The textual records of development programs in the SAVIX database are written in English, French and Portuguese. The texts in French and Portuguese were translated to English. IGA: income-generating activities.

Word	Frequency
education	55
IGA	46
health	44
financial	38
skills	38
training	34
gender	28
nutrition	27
business	22
development	20
food	19
protection	19

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). Fundamentals of predictive text mining. Springer.

Table 4.9: Cross-validation results to select the number of splitting variables in the models of random forests

n-vars	RMSE	MAPE	rho
1	27.6570	0.4743	0.5127
2	25.8645	0.4499	0.5496
3	25.2145	0.4409	0.5695
4	24.8961	0.4368	0.5789
5	24.7882	0.4362	0.5809
6	24.8208	0.4388	0.5786
7	24.9725	0.4440	0.5724
8	25.2004	0.4502	0.5640

n-vars: Splitting variables at each node
RMSE: Root mean squared error
MAPE: Mean absolute prediction error
rho: correlation between the estimations of ROS in the train sample and the predictions made by the model for the test sample

Zhou, Y., Tong, Y., Gu, R., & Gall, H. (2016). Combining text mining and data mining for bug report classification. Journal of Software: Evolution and Process, 28, 150–176.

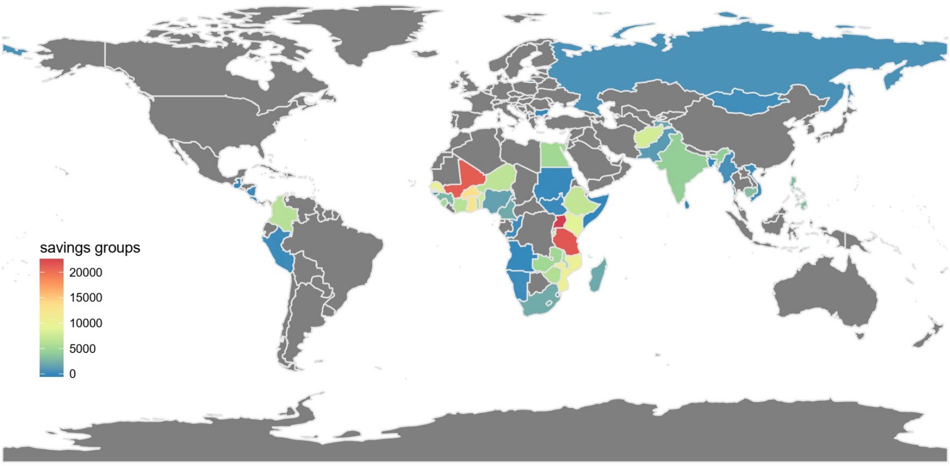


Figure 4.5: Geographical distribution of grass-root financial groups in the SAVIX database. Half of the groups in the database are located in eight African countries: Uganda (22702 groups), Tanzania (21374 groups), Mali (21021 groups), Burkina Faso (13680 groups), Ghana (12337 groups), Mozambique (10244 groups), Senegal (10148 groups) and Kenya (8906 groups).

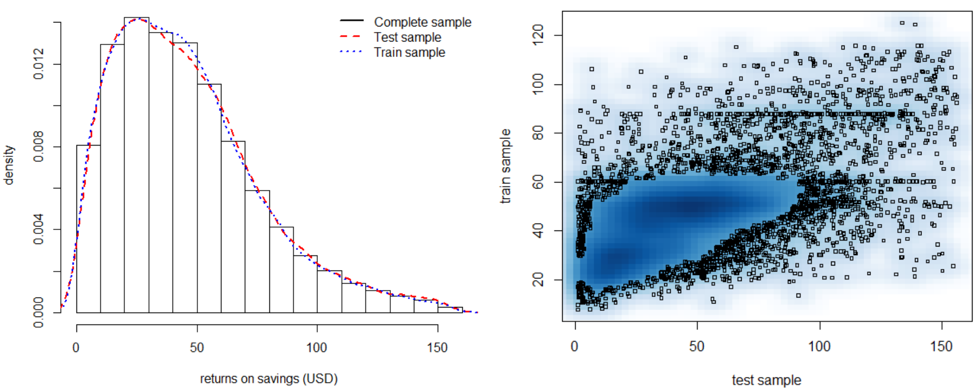


Figure 4.6: Left: Histogram of returns on savings (ROS) of grass-root groups in the SAVIX. Right: Scatterplot of the ROS in the trained and test samples used for cross-validation in the random forest.

Figure 4.7: Text-mining of the sustainable development programs provided to grass-root associations in the SAVIX. The size of the words represents the frequency of the records (higher frequency, higher size). As part of the text-mining exercise, the records of development programs in the SAVIX were transformed into a lower-case corpus and were cleaned from special characters, English stop-words, punctuations, extra white spaces and numbers.

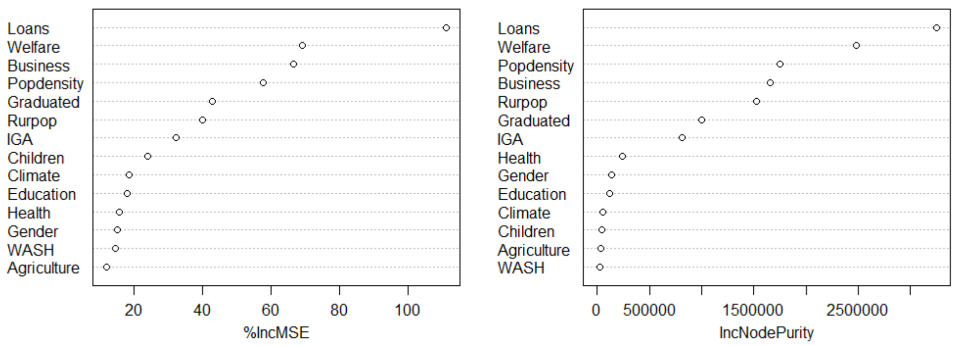


Figure 4.8: Results of the random forests. The importance of each variable for the generation of returns in grass-root financial associations is measured with the increase in the mean squared error of prediction (IncMSE, left) and the increase in node impurity (IncNodePurity, right). IncMSE is the error of prediction caused by a specific variable being excluded from the model. In the case of IncNodePurity, more relevant variables for prediction have a higher inter node variance and a smaller intra node variance.

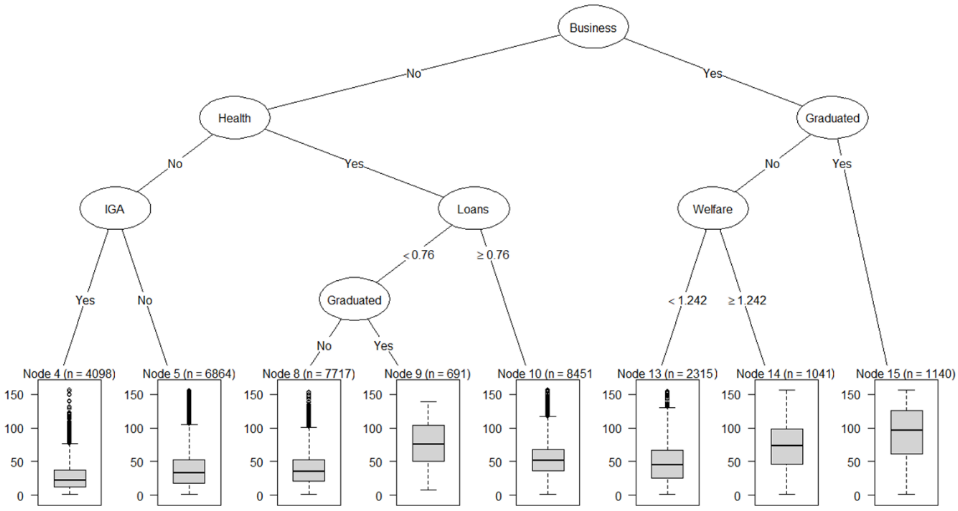


Figure 4.9: Decision tree of sustainable development programs offered to grass-root financial associations. In the graph below, n is the number of grass-root groups in each node of the tree.

D Paper D. Bifurcations in business profitability: An agent-based simulation of homophily in self-financing groups

Published in: Journal of Business Research (forthcoming).

Abstract

Formal financial institutions inadequately distribute startup capital to business ventures of ethnic minorities, women, low-educated, and young people. Self-financing groups fill this gap because in these associations agents accumulate their savings into a fund that is later used to provide loans to the members. This study builds and simulates an agent-based model that compares the profitability of businesses started by members of self-financing groups against businesses financed by commercial loans. The results indicate that—besides the self-generation of debt capital—businesses of members of self-financing groups can have higher returns due to the consolidation of social capital and the competitive advantage created through a dual process of homophily. Higher quotas of savings boost profits, but only up to a threshold, after which a bifurcation pattern—typical of complexity dynamics—emerges. The practical and theoretical implications of the findings are discussed and future research lines are proposed.

D.1 Introduction

Small businesses have a cumulative economic impact on the economy due to their relevance for long-term economic growth, productivity and job creation—see Acs and Armington (2006) or Haltiwanger et al. (2016). In Sub-Saharan Africa for example, micro-enterprises employ an estimated 80% of the working population Biekpe (2004), while in Colombia small business represent 96% of the enterprises created annually Santana (2017).

Despite their remarkable relevance, Cheng (2015) and Berger and Udell (2006) note that the financing of small startups is limited by the informational opacity that hinders banks from assessing the profitability, survivability and financial credibility of small ventures. Moreover, ethnic minorities, women, low-educated and young people are disproportionately impacted by the difficulties in accessing financial resources for business startups, even in developed countries[17].

The limitations in business financing can be overcome through government grants and subsidized loans. Due to the high cost of these policies, disadvantaged groups also rely on emerging financing instruments such as loan guarantees, microcredits, crowd-funding, peer-to-peer lending and business angel investment OECD (2014). One additional financing option for entrepreneurs who want to start their own business is self-financing groups.

Self-financing groups are a form of community-based associations that act as small savings and loan cooperatives of individuals Greaney et al. (2016). In a self-financing group, members agree to periodically provide an individual quota of savings with the aim of creating a collective pecuniary fund. The contribution is provided in group meetings during the life-cycle of a group. The fund of savings accumulated during the meetings is used for internal loan provision to the members.

Following Atlan (1991), self-financing groups can be conceptualized as a phenomenon of contextual complexity. Contextual complexity emerges from the communication among heterogeneous agents and the interaction of their goals. This pattern is typical of self-financing groups, because in these associations heterogeneous members with bounded rationality actively interact with each other over time, pursuing both personal and group-level goals, before and after a startup venture.

Due to its versatility, agent-based modeling is a computational approach suitable to capture the contextual complexity of self-financing groups. Members of self-financing groups face budget constraints and can exhibit random behavior, limiting the applicability of the traditional paradigm

[17] Smallbone et al. (2003) for example found that ethnic minorities in the UK are less successful in accessing bank loans and thus have to appeal to informal sources of startup finance. Cheng (2015) further remarks that this form of lending discrimination is even more pronounced with women and entrepreneurs from minority groups, who face a higher rate of loan denial and have unequal access to commercial credit from formal financial institutions. In the context of entrepreneurial activities in Europe, the OECD policy brief on access to business startup finance for inclusive entrepreneurship OECD (2014) further highlights that low-educated and young people are more likely to mention financing problems as a major constraint on starting a business.

of a representative, perfectly rational utility-maximizing agent (Farmer et al., 2005). In an agent-based model of autonomous and non-homogeneous agents, as the one described in Macal and North (2010), rational effects are dominated by stochastic fluctuations, and emerging social and financial patterns arise from the interaction between the behavioral and strategic decisions of heterogeneous agents with bounded rationality (Ponta & Cincotti, 2018).

In the agent-based model herein, agents decide to join a self-financing group due to intra-household conflicts, homophily and lack of access to formal financial loans. The members of the group are selected by an autonomous agent, who optimizes a gender-composition criterion (i.e., a preference for women members). Default rates are a function of the individual behavior of agents and the interaction among agents. Homophily enables the creation of social capital among members, which is aggregated to the debt capital generated by the group to start up a business. The sequential simulation of the algorithms produces a complex system in which patterns emerge from the interaction of agents at the micro level Pyka et al. (2018).

The results of the computational experiments in this study indicate that—due to homophily and embedded relational dynamics—funding from self-financing groups can increase business profitability compared to businesses financed through external loans, but only up to a bifurcation threshold. Self-financing groups build social capital that is difficult to imitate, which contributes as an additional resource to the success of a businesses initiative. However, after a threshold in the savings quota, a bifurcation in profitability emerges as a consequence of household budget constraints, the mimicking default behavior of agents, and the high interest rates that self-financing groups tend to charge for internal loans.

The findings of this study fill the research gap left by studies that have been traditionally focused on formal financing of small businesses. The impact of providing formal credit to small and micro business has been largely studied by *inter alia* Tuyon et al. (2011), Shahriar et al. (2016), Newman et al. (2017), Evelyn and Osifo (2018) or Atmadja et al. (2018)—see also the review of Chen et al. (2017). The conclusion of these studies is that formal financial institutions do not properly provide seeding to entrepreneurship, see for example Field et al. (2013) or Nguimkeu (2014). By contrast, there has been limited research on the impact of financing small business through internal loans from self-financing groups, creating a research gap that this study fills.

This study also contributes to the field of agent based modeling and complexity in business by using homophily to extend the recent literature on strategic group formation. In strategic group formation, agents maximize their individual utility by deciding to join or leave a social group Collins and Frydenlund (2018). Under a resource-view approach, the maximization of utility in self-financing groups translates to acquiring more resources in the form of loans and/or social capital. This study argues that homophily plays an additional ancillary role in strategic group formation and can further promote business profitability up to a bifurcation point.

The rest of the study is organized as follows: Section D.2 offers a conceptual overview of self-financing groups, agent-based modeling, social capital and homophily. Section D.3 describes the agent-based model of self-financing businesses. Section D.4 presents the results of simulating the model through computational experiments. Section D.5 concludes and discusses the practical and theoretical implications of the findings. A replication package with `MatLab` codes and step-by-step instructions to reproduce the results is also provided in an online supplementary material[18].

D.2 Conceptual framework

D.2.1 Self-financing groups

Self-financing groups are community-based organizations formed by people related by affinities or a specific goal Brody et al. (2015). The participants of a self-financing group join together to achieve individual and/or collective targets, which can be related to business startups, investment,

[18]The MatLab replication package is freely available at:
https://nl.mathworks.com/matlabcentral/fileexchange/73961-agent-based-model-of-nano-finance-groups

consumption-smoothing, asset acquisition or economic empowerment. The members of the group achieve their objectives through the accumulation of savings, the provision of internal credit and the creation of an informal insurance fund.

The roots of self-financing groups can be traced back to two types of indigenous associations: rotating savings and credit associations (ROSCAs) and accumulating savings and credit associations (ASCRAs). In ROSCAs, there is no loan provision because the pooled fund of savings is distributed to the members in a rotating pattern. In ASCRAs, the savings are not instantly redistributed but are rather accumulated in order to make loans with a fixed maturity Bouman (1995).

Self-financing groups are promoted by formal banks, the government or non-governmental organizations, who develop their own group schemes based on the principles of ROSCAs and ASCRAs. According to le Polain et al. (2018), the best-known facilitated self-financing models are the *village savings and loan association* initiated by CARE International, the *savings and internal lending communities* promoted by Catholic Relief Services and the *savings for changes* model promoted by Oxfam and Freedom from Hunger. In India, the National Bank for Agricultural and Rural Development (NABARD) steered the concept of self-help groups focused on the management of savings and credit Pillai and Abraham (2017).

Biggart (2001) relates the existence of self-financing groups to five situational circumstances: a communally-based social order, obligations that are held to be collective in nature, social and economic stability, social and economic isolation, and similarity between members. The research evidence has also discussed the importance of self-financing groups for capital accumulation Alila (1998), investment Hospes (1995), asset accumulation Annan et al. (2013), the promotion of income generating activities—Allen (2006), Ksoll et al. (2016) or Flynn and Sumberg (2018)—and the generation of social capital Ban et al. (2015).

Garikipati (2008) offers concrete examples about how self-financing groups can increase trust, which is the basis for social capital. For example, members of a self-financing group may help each other with childcare and livestock care without an explicit payment, or may help other members in finding waged work. Feigenberg et al. (2014) provide experimental evidence that shows that meetings of self-financing groups also aid to build social capital, measured by the number of times the members talk with each other about businesses.

D.2.2 Social capital and homophily

Loury et al. (1977) define social capital as naturally occurring social relationships aimed at promoting valued skills. Bourdieu and Wacquant (1992) understand social capital as resources accruing from a durable network, institutionalized through mutuality of acquaintance and recognition. Cooke and Wills (1999) make an additional distinction between human capital and social capital: while both refer to acquired skills, qualifications and capabilities, in social capital assets are less capable of formal certification.

Nahapiet and Ghoshal (1998) propose three facets of social capital: structural, relational and cognitive. The structural dimension refers to the degree of connectivity (the network) between agents. The relational dimension—which is based on trust and trustworthiness Fukuyama (1995), identity and identification Hakansson and Snehota (1995)—is based on the nature and characteristics of relationships, which can be competitive or cooperative. The cognitive dimension is a shared cognitive system of representations among agents, which can improve interpersonal communication and enhance relationships between members within an organization see, Jiang and Liu (2015, page 130).

Edwards and Foley (1997) raise two additional issues in the study of social capital: availability and equality. First, social capital is not equally available to all individuals, because geographic and social isolation limit the access to this resource. Second, not all social capital is created equal: the value of a specific source of social capital depends on the socioeconomic position of the individual within society. This inequality implies that agents will have heterogeneous levels of social capital depending on their socioeconomic and geographical characteristics Hsung et al.

(2017).

Theoretically, the importance of social capital for businesses can be seen from a resource-view approach if social capital is conceptualized as a source of competitive advantage that adds value to a venture, as in Jiang and Liu (2015). For instance, Bourdieu (1986) interprets social capital as an aggregate of actual or potential resources, again linked by a durable network of relationships, mutual acquaintances and recognition. N. Lin (2001) likewise suggests that social capital consists of resources embedded in social networks accessed and used by agents.

The resource-view approach to social capital has led Batjargal (2003) to propose that heterogeneity in the structural, relational and resource-based aspects of social capital is reflected in various aspects of business performance. The agent-based model of this study is based on the premise of Batjargal (2003): self-financing groups build social capital that is difficult to imitate and thus contribute as an additional resource to the success of a business initiative. Through a process of homophily, embedded relations—that improve coordination and reduce organizational conflict—influence purchase and sale decisions of entrepreneurs.

McPherson et al. (2001) define homophily as the principle that people tend to group with others who are like them. According to Collins and Frydenlund (2018), the factors that determine homophily include gender, religion, social class, education and other intrapersonal or behavioral characteristics. Granovetter (1985) and Jiang and Liu (2015) argue that intrapersonal and behavioral characteristics of members of a group create dense ties that support stronger reciprocity and greater trust. Social capital emerges from these dense ties, which minimize monitoring and transaction costs by reducing opportunistic behaviors Uzzi (1997).

D.2.3 Agent-based modeling

Agent-based modeling can be traced back to the developments of complexity theory and artificial intelligence—see Weisbuch (1991), Kauffman (1993), Order (1995), Langton (1997) or Macal and North (2010). Complexity analyzes patterns and structures that emerge from interactions Kirman (2010). Artificial intelligence, in turn, is a subfield of computer science aimed at building agents that exhibit aspects of intelligent behavior in terms of autonomy, social ability, reactivity and proactiveness Wooldridge and Jennings (1995). Based on the interactions among intelligent agents, agent-based models produce insights that guide decision-making, help to solve complex problems and simulate real-life phenomena.

Hamill and Gilbert (2016) define an agent-based model as a computer program that creates an artificial world of heterogeneous agents and enables the investigation of their interactions. In this artificial world, agents react to other agents, pursue goals, communicate with other agents and move around within the environment Wooldridge and Jennings (1995).

According to Macal and North (2010), an agent-based model has three elements:

(i) A set of agents, with attributes and behaviors.

(ii) A set of agent relationships and methods of interaction, i.e. a topology of connectedness that defines how and with whom agents interact.

(iii) The environment. Besides interacting with other agents, agents can in some cases also affect their environment.

Following Lemos (2017), an agent $\mathscr{A}$ can be defined as a computer system that is situated in some environment and is capable of perceiving, deciding and performing actions in an autonomous way. Formally, let $\mathsf{E} \in \mathbb{Z}^+$ be the set of possible environment states, and let $\mathsf{A} \in \mathbb{Z}^+$ be the set of actions available to $\mathscr{A}$, then the sequence of environment states alternating with actions of $\mathscr{A}$ can be defined using the run of simulations $\mathscr{R}$, where $\mathscr{R}^{\mathsf{A}} \subset \mathscr{R}$ is the subset of runs ending with an action, and $\mathscr{R}^{\mathsf{E}} \subset \mathscr{R}$ the subset of runs ending with an environment state. Based on the definitions above, an agent will be a function that maps runs ending in environment states into actions: $\mathscr{A} : \mathscr{R}^{\mathsf{E}} \mapsto \mathsf{A}$ see, Wooldridge (2009).

An agent-based topology defines how agents are connected to each other. Typical topologies are cellular automata Wolfram (2018), the Euclidean space, networks where nodes are the agents and the links are relationships El-Sayed et al. (2012), spatial grids—based on a geographic information system (GIS)—and aspatial topologies where agents have no location because it is not relevant for the simulation at hand.

In some models, agents can also affect and modify their environment when the collective action of multiple agents causes changes in the environmental state in which agents operate, thereby generating the map $A \mapsto \mathcal{R}^E$. See *inter alia* Sengupta and Bennett (2003), who use a model of agents distributed in a geographical environment to simulate the ecological and economic impacts of agricultural policies.

Emerging patterns are also a characteristic of agent-based models. As Macal and North (2010) highlight, both the heterogeneity of agents and self-organization are features of agent-based simulation that allow the emergence of complexity patterns. This emergence differentiates agent-based models from other simulation techniques, such as discrete-event simulation and system dynamics.

D.3 Agent-based model of self-financing businesses

Small businesses play an important role in economic growth and socioeconomic development Tuyon et al. (2011). Startup businesses normally confront a shortage of capital and limited access to loans from formal commercial banks and thus have to draw upon informal sources of startup finance, such as micro-credits provided by formal finance institutions or internal loans obtained from informal self-financing groups.

The agent-based model of this study aims to simulate the profitability of businesses financed by self-financing groups. The business profitability of the self-financing group is compared to the profitability of a counterfactual business financed with external loans from a formal financial institution.

In the agent-based simulation of self-financing groups (henceforth, ABS-SFG):

— The set of active agents are (i) women and men of the working population in an artificial community, and (ii) an autonomous agent in charge of creating the self-financing group. Passive agents are children and the elderly in the community, who do not make decisions but influence the behavior of active agents.

— The topology is defined in the Euclidean space $\mathbb{R}^n$, i.e. the set of all real n-tuples $\mathbb{R}^n = \{(p_1, ..., p_n)|p_j\}$ for a real number p_j in $j = 1, 2, ..., n$ Abbena et al. (2017). The connectedness of the agent-based model in the Euclidean two-dimensional space $\mathbb{R}^2$ is calculated using (i) the Euclidean distance between households and (ii) the Euclidean distance between the intrinsic demographic characteristics of individuals.

— The environment is defined only by the interaction of agents with other agents. Agents cannot change their environment.

The computational ABS-SFG model is a multilayered simulation of four algorithms that run sequentially in two phases (Figure 4.10). The agent-based simulation illustrated in Figure 4.10 is a 'microverse' containing the dynamics and environment of self-financing groups, as in Guterman et al. (2015). The model creates an artificial world that emulates the behavior of the members of self-financing groups in a village, as in the second-order *simulacra* of Baudrillard (1994), who inspired the simulated reality of Wachowski and Wachowski (1999).

The four stages of the ABS-SFG model can be grouped into an initialization phase (algorithms 1 and 2) and a running phase (algorithms 3 and 4). In the first stage, the model starts simulating a community of agents in an artificial village (algorithm 1). In the second stage, a self-financing group is formed by an agent that selects members from the individuals in the artificial village who want to be part of the group (algorithm 2). In the third stage, heterogeneous agents in the self-financing group interact with each other to accumulate social and debt capital (algorithm 3). In the last stage, an internal loan is provided to agents for the creation of a business venture and the profitability of a self-financing business is compared with the profitability of a counterfactual business of non-members financed by a loan from a formal financial institution (algorithm 4).

Table 4.10 shows the submodels in the algorithms and lists the variables/traits included in each submodel. Table 4.10 also indicates which equations are used to calculate the traits in each submodel and further clarifies whether values are predetermined or produced by the model. The next subsections describe in detail the equations and submodels in each algorithm.

D.3.1 Algorithm 1: Artificial community

Box 1 shows the first algorithm of the simulation model. Based on the number of households (H), random numbers from probability distributions are used to create an artificial community of agents that have three demographic characteristics: age (a), gender (g), and the number of dependent individuals in the household (i.e. children and the elderly, δ_h).

The h-households in the village ($h = 1, 2, ..., H$) are populated with i-individuals based on the numerical values of a centered probability mass function generated from a discrete Poisson distribution:

$$i_h(\lambda_h) = \lambda_h + \exp^{-\lambda_h} \frac{\lambda_h^{i_h}}{i_h!}, \tag{4.1}$$

The stochastic function in equation 4.1 was chosen to populate the households following V. Jennings et al. (1999) and V. Jennings and Lloyd-Smith (2015), who show that a Poisson process is suitable for modeling household size distribution. The number of productive individuals in the household (δ_h) is obtained from random numbers of a discrete uniform distribution (equation 4.2), while the gender of each individual (g_i) is obtained from numerical values of a conditional uniform discrete distribution (equation 4.3):

$$\delta_h \sim \mathcal{U}(1, u_\delta) \tag{4.2}$$

$$g_i | \delta_h \sim \mathcal{U}(1, u_{\delta,g}), \quad u_{\delta,g} = 2u_\delta. \tag{4.3}$$

The age of each i-individual (a_i) is produced from a mixture of discrete uniform distributions:

$$\begin{aligned} a_i &\sim \mathcal{MU}(\alpha_a) \\ &= \sum_{j \in \mathcal{Z}^{1,2,3}} \mathcal{U}_{ij}(1, u_{age}) + \sum_{j \in \mathcal{Z}^{4,5}} \mathcal{U}_{ij}\left(1, \frac{1}{2}u_{age}\right). \end{aligned} \tag{4.4}$$

The income (y_i) of the working population in the village is generated using random numbers from a log-normal distribution:

$$f(y_i | \mu_g, \sigma_g) = \frac{1}{y_i \sigma_g \sqrt{2\pi}} \exp^{\frac{(\ln y_i - \mu_g)^2}{2\sigma_g^2}}, \tag{4.5}$$

where $g \in \{w, \neg w\}$ is a gender index for women (w) and men ($\neg w$), under the assumption that men in the population have (on average) higher income than women ($\mu_{\neg w} > \mu_w$) and less dispersion around the average income ($\sigma_{\neg w} < \sigma_w$). The stochastic function for income was chosen as log-normal because although income follows a Pareto law in the upper tail, the distribution of the low-income population is normally described with a log-normal distribution; see for example Souma (2001) or Banerjee et al. (2006). The assumption about the difference of the distribution of income for women is based on evidence about the polarization of women's employment and income, which has been related to occupational segregation, discrimination, work-life balance, part-time work, career patterns across the life cycle and labor mobility—see Hakim (2016).

The first stage of the simulation produces a population matrix $\mathbf{P}_{\mathcal{A}_i}$ with the following agent's characteristics: household location (h) of agents, identification of individuals (i_h) in the population, gender (g_i), number of productive individuals in a household (δ_h), age of the agents (a_i) and income (y_i). See Box 1 below.

Box 1. Algorithm 1: Artificial community

Data: $H, \lambda_h, u_\delta, u_{age}, \mu_g, \sigma_g$
Result: $\mathbf{P}_{\mathcal{A}_i} \ni \{h, i_h, \delta_h, g_i, a_i, y_i\}$
for $H \leftarrow h$ **do**

$\quad i_h(\lambda_h) = \lambda_h + \exp^{-\lambda_h} \frac{\lambda_h^{i_h}}{i_h!}$

$\quad \delta_h \sim \mathcal{U}(1, u_\delta)$

$\quad$ **for** $i_h \leftarrow i$ **do**

$\quad\quad g_i|\delta_h \sim \mathcal{U}(1, u_{\delta,g}), \quad u_{\delta,g} = 2u_\delta$

$\quad\quad a_i = \sum_{j\in\mathcal{Z}^{1,2,3}} \mathcal{U}_{ij}(1, u_{age}) + \sum_{j\in\mathcal{Z}^{4,5}} \mathcal{U}_{ij}\left(1, \frac{1}{2}u_{age}\right)$

$\quad\quad f(y_i|\mu_g, \sigma_g) = \frac{1}{y_i\sigma_g\sqrt{2\pi}} \exp\left(\frac{(\ln y_i - \mu_g)^2}{2\sigma_g^2}\right)$

$\quad$ **end**

end

D.3.2 Algorithm 2: Formation of a self-financing group

In the second algorithm, an autonomous field agent $\mathcal{A}_f$ creates a self-financing group by selecting members from the subset of the individuals $P_{\mathcal{A}_i}$ of the population $\mathbf{P}_{\mathcal{A}_i}$ who want to join the group $(P_{\mathcal{A}_i} \subset \mathbf{P}_{\mathcal{A}_i})$. Due to the probabilistic nature of the agent's wish to join a SFG—and due to the optimization decision of the autonomous field agent when deciding on gender composition—the m-number of members of a self-financing group is not programmed in the model, but is rather one of the emerging patterns produced by the model.

Following the theory of strategic group formation Collins and Frydenlund (2018), agents join or leave a group in order to gather social and financial resources. Besides this utility maximizing behavior, homophily plays a role in the formation of self-financing groups. In the ABS-SFG model, the probability $\mathbb{P}_i(m)$ of i-agents wishing to join a self-financing group is a quadratic mixture of the probabilities related to their lack of financial access (f_i^a), the geographical proximity among households in the village (ψ_i^h), the social connections among productive individuals (s_i), and intra-household conflicts (h_i^c):

$$
\begin{cases}
\mathbb{P}_i(m) = \omega_{\mathbb{P}_i(m)}^2 \mathbb{P}(f_i^a) + \left(1 - \omega_{\mathbb{P}_i(m)}\right)^2 \left(\mathbb{P}(\psi_i^h) + \mathbb{P}(s_c) + \mathbb{P}(h_i^c)\right) & (4.6)\\
\mathbb{P}(f_i^a) = 1 - \dfrac{1}{1 + e^{1 - y_{h,i}}} & (4.7)\\
\mathbb{P}(\psi_i^h) = \dfrac{1}{1 + e^{1 - (\sum_i^N \sqrt{(h_i - h)^2})^{-1}}} & (4.8)\\
\mathbb{P}(s_i) = \dfrac{1}{1 + e^{1 - i_{h,p}}} & (4.9)\\
\mathbb{P}(h_i^c) = \dfrac{1}{1 + e^{1 - i_{h,\neg p}}} & (4.10)
\end{cases}
$$

In equation 4.6, $\omega_{\mathbb{P}_i(m)}$ is the weight—the importance—that individuals assign to lack of formal financial access, $\mathbb{P}(f_i^a)$. Lack of financial access in equation 4.7 is based on the probabilistic transformation of the income of each individual in a household $(y_{h,i})$. Demirgüç-Kunt et al. (2008) argue that cost-effective micro-financial services are not available to the extreme poor due to the imbalance between the fixed transactions costs of formal financial institutions and the small transactions and low demand of the extreme poor, which cannot be compensated with higher interest rates. Stiglitz and Weiss (1981) add that, in the presence of imperfect and costly information, the expected rate of return of banks increases less rapidly than the interest rate and, beyond a point, may actually decrease, thus generating a credit-rationing effect in formal banking. Hence, the low income of the agents in a village reduces the probability of having access to a formal loan from a financial institution.

Homophily is implemented through the sigmoid functions 4.8 and 4.9, which transform to probabilities the geographical proximity of households (equation 4.8), as well as the connections among productive individuals (equation 4.9). Homophily plays a dual role in self-financing groups: during group formation and during the life-cycle of the group. During group formation, homophily interacts with the utility maximization behavior of agents who seek resource acquisition, because self-financing groups are generally formed by peers who share similar socioeconomic and demographic characteristics. During the life-cycle of the group, homophily consolidates social capital and reduces the risk aversion among agents.

In order to measure the probability of joining a group based on geographical homophily, the Euclidean distance between agent's households (d_h) is given by $\|d_i\| = (\sum_i^N \sqrt{(h_i - h)^2})^{-1}$, and this distance is converted to a probability measure through the sigmoid function of equation 4.8, which assigns more homophily to individuals living in households near the center of the village. The probability of joining a group due to social homophily in equation 4.9 is based on the probabilistic transformation of the number of productive individuals in a household $(i_{h,p})$. Households with a large number of productive individuals have more social connections with other productive agents, and thus have a higher probability of joining a self-financing group.

Finally, intra-household conflicts (equation 4.10) are measured by the number of dependents in a household $(i_{h,\neg p})$, since a large number of dependent children and retired elderly can lead to higher intra-household conflicts among productive members in relation to investment decisions, and thus can increase the probability of agents joining self-financing groups. Conflictual interactions within a household has been put forward by Anderson and Baland (2002) as one of the main reasons to join self-financed groups when there are asymmetric preferences between men and women about investment in household goods.

The set of members that want to be part of the self-financing group $(P_{\mathcal{A}_i})$ is obtained with a rejection sampling algorithm in which the candidates are agents $\mathbf{P}_{\mathcal{A}_i}$ for which the mixture probability $\mathbb{P}_i(m)$ in equation 4.6 is higher than a random number $u_m \sim \mathcal{U}(0,1)$, where $\mathcal{U}(\cdot)$ is a standard continuous uniform distribution.

The autonomous field agent $\mathcal{A}_f$ selects the $\mathbb{M}_{\mathcal{A}_i}$ members of the self-financing group from the set of individuals that want to be part of the group, $\mathbb{M}_{\mathcal{A}_i} \subset P_{\mathcal{A}_i}$ $(P_{\mathcal{A}_i} \subset \mathbf{P}_{\mathcal{A}_i})$. The autonomous field agent $\mathcal{A}_f$ that forms a group is commonly called 'field officer' by development agencies. Self-financing groups are promoted by development agencies that hire and pay an agent—the field officer—to create, train and supervise a group; see Allen and Panetta (2010).

The criterion of a field officer $\mathcal{A}_f$ for selecting the members $x_i \in \mathbb{M}_{\mathcal{A}_i}$, $i = 1, 2, ..., m$, is to have more women than men in the group. This positive gender discrimination is related to the fact that facilitating agencies—which pay and instruct the field officer $\mathcal{A}_f$—tend to target women because they consider women to make a higher contribution to family welfare, since women give priority to spending their earnings on their children Guha and Gupta (2005). Rasmussen (2012) also attributes the gender focus of self-financing groups to women's economic resilience, since savings enable women to handle income shocks and confront unforeseen emergencies such as illness or loss of employment Ghosh and Vinod (2017).

Formally, when selecting the members $\{x_1, x_2, ..., x_m\} \in \mathbb{M}_{\mathcal{A}_i}$ from the potential set of candidates $P_{\mathcal{A}_i}$, an artificial agent $\mathcal{A}_f$ wants to achieve a gender ratio of women to men τ higher than

$\tau_w \in (0, 1]$,

$$S_{\mathcal{A}_f}(P_{\mathcal{A}_i}, \tau) = \begin{cases} x_1, x_2, ..., x_m & \text{if } \tau \geqslant \tau_w \\ \emptyset & \text{else} \end{cases} \tag{4.11}$$

Equation 4.11 is computationally equivalent to a *while* loop. See the algorithm in Box 2 below.

Box 2. Algorithm 2: Formation of a self-financing group (group creation and members' selection)

Data: $\mathbf{P}_{\mathcal{A}_i} \ni \{h, i_h, g_i, \delta_h, a_i, y_i\}$
Result: $\mathbb{M}_{\mathcal{A}_i} \subset P_{\mathcal{A}_i} \subset \mathbf{P}_{\mathcal{A}_i}$
$\mathbb{P}_i(m) = \omega^2_{\mathbb{P}_i(m)} \mathbb{P}(f^a_i) + \left(1 - \omega_{\mathbb{P}_i(m)}\right)^2 \left(\mathbb{P}(\psi^h_i) + \mathbb{P}(s_i) + \mathbb{P}(h^c_i)\right)$
$\mathbb{P}(f^a_i) = 1 - \frac{1}{1+e^{1-y_{h,i}}}$
$\mathbb{P}(\psi^h_i) = \frac{1}{1+e^{1-(\sum_i^N \sqrt{(h_i-h)^2})-1}}$
$\mathbb{P}(s_c) = \frac{1}{1+e^{1-i_{h,p}}}$
$\mathbb{P}(h^c_i) = \frac{1}{1+e^{1-i_{h,\neg p}}}$
$u_m \sim \mathcal{U}(0, 1)$
if $\mathbb{P}_i(m) > u_m$ **then**
  $i_h \in P_{\mathcal{A}_i}$
else
  $i_h \notin P_{\mathcal{A}_i}$
end
while $\tau < \tau_w$ **do**
  $S_{\mathcal{A}_f}(P_{\mathcal{A}_i}, \tau)$
  if $\tau \geqslant \tau_w$ **then**
    $\{x_1, x_2, ..., x_m\} \ni \mathbb{M}_{\mathcal{A}_i}$
  else
    $\emptyset$
  end
end

D.3.3 Algorithm 3: Agent-based simulation of self-financing groups and formation of social capital

Algorithm 3 simulates the dynamics of savings accumulation as well as the formation of social capital among the members of a self-financing group. In the agent-based simulation, the emerging patterns of members' default and fund accumulation are the result of (*i*) the interactions among members and (*ii*) an adaptive rule—a rule that changes the rules—for savings accumulation, which is activated when a large number of members do not contribute with their savings to the group.

In a self-financing group, each $i = 1, 2, ..., m$-member contributes a ρ amount of savings to a common fund during the life-cycle of the group (Burlando & Canidio, 2017). This life-cycle is a round of meetings $1, 2, ..., t$ where the members get together to contribute their quota of savings to the common fund. In the ABS-SFG model, the number of meetings t is equal to the number of members m in a group, to account for the fact that larger groups need longer organizational periods.

The number of members of a group (m) is not predetermined but is rather an emergent parameter produced by the interactions of agents in the model. The m number of members of the simulated groups is similar to the number of members observed in real life: around 20 members. Bisrat et al. (2012) notes that SFGs have this number of members because—although more members allow to accumulate a larger sum of money over a cycle—too many members involve a greater number of administrative problems, thus creating an incentive to keep the number of participants to around 20 members.

A member of a self-financed group enters a state of default *in savings* if the agent does not contribute his/her quota of savings during a meeting t. In the model, this default state is modeled as an inequality between the idiosyncratic probability of default $\mathbb{P}_i(d_s)$ and the group-level extrinsic probability of default $\mathbb{P}_e(d_s, t)$:

$$\mathbb{P}_i(d_s) > \mathbb{P}_e(d_s, t) + u, \quad u \sim \mathcal{U}(0, 1), \tag{4.12}$$

where u is a random component from a uniform distribution $\mathcal{U}(0, 1)$ that models the unexpected events that can increase the probability of default in savings.

The idiosyncratic probability of default $\mathbb{P}_i(d_s)$ is a convex combination of each member's intrinsic probability of default, related to age (a_i) and income (y_i), and weighted by gender (γ_g):

$$\begin{cases} \mathbb{P}_i(d_s) = \gamma_g \left(\omega_{d_s} \mathbb{P}_{d_s}(a_i) + (1 - \omega_{d_s}) \mathbb{P}_{d_s}(y_i)\right) & (4.13) \\ \mathbb{P}_{d_s}(a_i) = 1 - \dfrac{1}{1 + e^{1 - (\sum_i^m \sqrt{(a_i - a)^2})^{-1}}} & (4.14) \\ \mathbb{P}_{d_s}(y_i) = 1 - \dfrac{1}{1 + e^{1 - y_i}} & (4.15) \end{cases}$$

where $\gamma_g \in \mathbb{R}^{0,1}$ and $\omega_{d_s} \in \mathbb{R}^{0,1}$ are predetermined parameters, $\mathbb{P}_{d_s}(a_i)$ is the probability of default related to the age of an agent, and $\mathbb{P}_{d_s}(a_i)$ is the probability of default related to the income of an agent.

In equation 4.13, the parameter $\gamma_g \in \mathbb{R}^{0,1}$ measures higher female repayment rates when $\gamma_g \to 1$, as reported in, for example, Mayoux et al. (2000) or Gonzales Martınez et al. (2019). The probability of default related to the age of an agent $\mathbb{P}_{d_s}(a_i)$ in equation 4.14 is calculated using the inverse Euclidean distance from the centroid of the age in the group. The parameter ω_{d_s} is the weight (the importance) of age for the probability of default in savings. $\mathbb{P}_{d_s}(y_i)$ is the probability of default related to the income of an agent. Individuals with low income, compared to the rest of the members, have a higher probability of entering a state of default in savings. Likewise, individuals in the tails of the age distribution (young and older members) have a higher probability of default, compared to other members. Gender is included as an interaction term, assuming that women are financially more reliable and thus have a lower probability of entering a default state compared to men—see Abbink et al. (2006) or D'espallier et al. (2011).

The extrinsic probability of default $\mathbb{P}_e(d_s, t)$ depends on group-level characteristics that arise from the interaction among members, the amount of savings contribution and the stage of meetings in the life-cycle of the group:

$$\begin{cases} \mathbb{P}_e(d_s, t) := \tanh z_e \equiv \frac{\sinh z_e}{\cosh z_e} = \dfrac{e^{z_e} - e^{-z_e}}{e^{z_e} + e^{-z_e}} \\ z_e = 1 - \beta\left(\rho + \sqrt{m} - t\right). \end{cases} \tag{4.16}$$

In equation 4.16, ρ is the individual amount of savings that each agent has to contribute to the common fund. Higher amounts imply a higher burden for the individuals and thus increase the probability of default. Parameter t is added to the default threshold z_e to reflect the fact that the probability of entering a default state increases over time. Conversely, $\sqrt{m}$ reduces the probability of default in larger groups, because peer pressure in such groups can act as a savings commitment device. For example, Kast et al. (2012) conducted a randomized trial with microentrepreneurs in

Chile and found that peers in savings groups provide a mutual service by regularly holding each other accountable for setting savings goals and regularly reminding each other of these goals.

Agents that enter into a state of default in a meeting $t - 1$ will also affect the behavior of other agents in the next meeting t, because non-defaulting agents that mimic the behavior of defaulting agents will fail to deliver their quota of savings. This mimicking behavior is modeled in equation 4.16 through a switching parameter $\beta \in \{\beta_d, \beta_{\neg d}\}$ that changes when agents enter a default state d in the group ($\beta_d > \beta_{\neg d}$). The predetermined parameter of mimicking behavior β increases the chances that the rest of the non-defaulting members will enter a default state when another member fails to deliver his/her quota of savings. Larger values of β imply that a defaulting agent can strongly interact and dramatically affect the behavior of the rest of the agents in a self-financing group.

The pattern of savings accumulation in the common fund (b) of the self-financed group is defined by,

$$b := \sum_{t=1}^{T} b_t = \begin{cases} \sum_{t=1}^{m} (m - d_t)\rho & \text{if } \dfrac{d_t}{m} \leqslant \tau_d, & (4.17) \\ \sum_{t=1}^{m} (m - d_t)\rho(1 + u) & \text{if } \dfrac{d_t}{m} > \tau_d, & (4.18) \end{cases}$$

It is common in agent-based models to introduce *adaptation*, where agents learn or adapt by changing their rules and behavior based on their experience and dynamic interactions Smith and Conrey (2007). In the case of a self-financing group, a high default rate in savings can dramatically reduce the fund accumulated for loans. Thus, to compensate for this reduction, an adaptive rule ('a rule that changes the rules') is introduced in the agent-based model (equation 4.18): groups with a high default rate of savings ($\frac{d_t}{m} > \tau_d$) change the pattern of fund accumulation from a fixed scheme to a solidarity scheme, in which non-defaulting members provide an additional contribution—$\rho(1 + u), u \sim \mathcal{U}(0, 1)$—beyond the quota ($\rho$) due at each meeting t, in order to stabilize the collective savings fund b_t over time. See the algorithm in Box 3.

Box 3. Algorithm 3: Agent-based simulation of self-financing groups and formation of social capital

---

Data: $\mathbb{M}_{\mathcal{A}_i}, \rho$

Result: $k_m \ni \{k_l, k_s\}$

$\mathbb{P}_i(d_s) = \gamma_g \left(\omega_{d_s}\mathbb{P}_{d_s}(a_i) + (1-\omega_{d_s})\mathbb{P}_{d_s}(y_i)\right)$

$\mathbb{P}_{d_s}(a_i) = 1 - \frac{1}{1+e^{1-(\sum_i^m \sqrt{(a_i-a)^2})^{-1}}}$

$\mathbb{P}_{d_s}(y_i) = 1 - \frac{1}{1+e^{1-y_i}}$

for $meeting \leftarrow t$ **do**

  $u \sim \mathcal{U}(0,1)$

  $\beta \in \{\beta_d, \beta_{\neg d}\}$

  $\mathbb{P}_e(d_s, t) := \tanh z_e = \dfrac{e^{z_e} - e^{-z_e}}{e^{z_e} + e^{-z_e}}$

  $z_e = 1 - \beta\left(\rho + \sqrt{m} - t\right)$

  if $d_{\forall i}(t-1) = 1$ **then**

    if $\mathbb{P}_i(d_s) > \mathbb{P}_e(d_s, t) + u$ **then**

      $d_i(t) = 1$

    else

      $d_i(t) = 0$

    end

  else

    $\beta_{\neg d} < \beta_d$

    if $\mathbb{P}_i(d_s) > \mathbb{P}_e(d_s, t) + u$ **then**

      $d_i(t) = 1$

    else

      $d_i(t) = 0$

    end

  end

  if $\frac{d_t}{m} \leqslant \tau_d$ **then**

    $\sum_{t=1}^{T} b_t = \sum_{t=1}^{m}(m - d_t)$

  else

    $\sum_{t=1}^{T} b_t = \sum_{t=1}^{m}(m - d_t)\rho(1+u)$

  end

end

$k_l \equiv \ell$

$k_s = \gamma_s\left(\|a_i\| + \|y_i\| + \|h_i\|\right)$

$\|a_i\| = \left(\sum_{i=1}^{m}(a_i - a)^2\right)^{-1/2}$

$\|y_i\| = \left(\sum_{i=1}^{m}(y_i - y)^2\right)^{-1/2}$

$\|h_i\| = \left(\sum_{i=1}^{m}(h_i - h)^2\right)^{1/2}$

The agent-based algorithm in Box 3 produces two outputs: debt capital (k_ℓ) and social capital (k_s). Debt capital is a fraction of the accumulated fund b and is discussed in section D.3.4. The formation of social capital, in turn, is the result of the homophily among the participants of a self-financing group.

The quantitative operationalization of social capital in the model is based on the multilevel ecometric approach of Raudenbush and Sampson (1999). This approach allows one to differentiate between individual and area-level sources of variation in social capital Mackenbach et al. (2016). Formally, social capital is calculated as a function of the Euclidean distance between individuals in an artificial community, in terms of the homophily related to their age ($\|a_i\|$), income ($\|y_i\|$) and household location ($\|h_i\|$):

$$
\begin{cases}
k_s = \gamma_s \left(\|a_i\| + \|y_i\| + \|h_i\|\right) & (4.19) \\
\gamma_s = 1 + \dfrac{d}{m} & (4.20) \\
\|a_i\| := \left(\sqrt{\sum_{i=1}^{m}(a_i - a)^2}\right)^{-1} = \left(\sum_{i=1}^{m}(a_i - a)^2\right)^{-1/2} & (4.21) \\
\|y_i\| := \left(\sqrt{\sum_{i=1}^{m}(y_i - y)^2}\right)^{-1} = \left(\sum_{i=1}^{m}(y_i - y)^2\right)^{-1/2} & (4.22) \\
\|h_i\| := \sqrt{\sum_{i=1}^{m}(h_i - h)^2} = \left(\sum_{i=1}^{m}(h_i - h)^2\right)^{1/2}. & (4.23)
\end{cases}
$$

A multiplying gamma factor (γ_s) is included in the formula of social capital of equation 4.19 to account for the impact of being part of a self-financing group. The parameter γ_s accounts for the fact that non-default members of a self-financing group create additional bonds of trust, trustworthiness and reciprocity. These resources, according to Putnam (1993), promote their ability to undertake collective actions, such as starting a joint business venture.

In the case of agents that are not part of a self-financing group, their social capital is a function of their own homophily (i.e., $\gamma_s = 1$ for $\neg M_{\mathcal{A}}$), while in the case of non-defaulting members of a self-financing group $\gamma_s > 1$ because the ratio of defaulting members to the total members of a group is added to the scale parameter of social capital (equation 4.20). This implies that in a group with a large number of defaulting members (d), stronger social bonds of trust will be created among the remaining non-defaulting members $m - d$.

In equation 4.21, less social cohesion is assigned to individuals that are not closer to the average age of the population (a), on the basis of studies of reduced social capital in young and elderly populations (Lauder et al., 2006). Lower social bonds are also allocated to individuals that have an income in the tails of the distribution (equation 4.22), since income inequality has been found to be related to a reduction in social cohesion; see Khambule and Siswana (2017). Finally, less contextual social capital is assigned to individuals that live in households located in the village periphery (equation 4.23). This last area-level allocation of social capital is based on the literature on neighborhood formation of social capital; see Butler and Robson (2001) and Forrest and Kearns (2001).

D.3.4 Algorithm 4: Loan allocation and business simulation

Box 4 shows the last stage of the simulation (algorithm 4). In the last algorithm, agents start a joint business venture with the social capital and the debt capital obtained after being part of a self-financing group. It is assumed that the group members start a business together; this a direct result of homophily and tends to be common in low-income groups of women, as those served by self-financing groups. See, for example, the cases of informal businesses in Africa described in Spring (2009).

The probability of members receiving a loan (ℓ) from the self-financing group is conditional on the income (y_i) and social capital ($k_{s,i}$) of an i-individual:

$$\begin{cases} \mathbb{P}(\ell|y_i, k_s) := \tanh z_\ell \equiv \frac{\sinh z_\ell}{\cosh z_\ell} = \frac{e^{z_\ell} - e^{-z_\ell}}{e^{z_\ell} + e^{-z_\ell}} \\ z_\ell = \frac{2 + e\left(e^{-y_i} + e^{-k_{s,i}}\right)}{(1 + e^{1-y_i})(1 + e^{1-k_{s,i}})}. \end{cases} \tag{4.24}$$

A member of the SFG will receive a loan if his/her probability of receiving a loan is higher than the ex-post risk aversion of the self-financing group (r_{post}):

$$\begin{aligned} \mathbb{P}(\ell|y_i, k_s) &> r_{post} \\ &> r - f(\omega_r; k_{s,(m-d)}) \end{aligned} \tag{4.25}$$

In equation 4.25, $r \in \mathbb{R}^{0,1}$ is the ex-ante risk aversion of the SFG. This is the risk aversion toward providing loans at the start of the group meetings. This risk is updated by non-defaulting members after being part of a SFG, on the basis of a Gompertz function of social capital ($k_{s,(m-d)}$),

$$f(\omega_r; k_{s,(m-d)}) = \omega_r e^{-\xi \omega_r e^{-\frac{\omega_r k_s}{\xi (m-d)}}}.$$

In the function $f(\omega_r; k_{s,(m-d)})$, ξ is the standard scientific notation $\xi_m \times 10^{\xi_n}$, for which a mantissa and an order of magnitude of $\xi_m = \xi_n = 1$ generate a smoothed curve saturated toward the asymptote ω_r (Laird, 1964); this is, the social capital of the non-defaulting agents $k_{s,(m-d)}$ reduces the ex-ante risk aversion of the SFG only up to an asymptotic ω_r-probability:

$$\lim_{k_{s,(m-d)} \to \infty^+} f(\omega_r; k_{s,(m-d)}) = \omega_r.$$

A value of 0.5 was chosen for the asymptotic risk-reduction probability ω_r (the hyperparameter $\omega_r = 0.5$), based on Laplace's uncertainty principle: if no additional information about the reliability of the potential borrowers is available, in the limit the SFG members assume that all possible events are equiprobable; see *inter alia* Gurov (2005).

The amount of the loan allocated to the borrowers (ℓ) is a fraction of the total savings in the common fund (b),

$$k_\ell := \ell = b(1 - r_{post}), \tag{4.26}$$

with an effective interest rate equal to the nominal interest rate plus the updated (ex-post) risk aversion of the self-financing group,

$$i_\ell = i + r_{post}. \tag{4.27}$$

Equations 4.25, 4.26, and 4.27 capture the pattern of loan provision and bucketization of interest rates—as a function of risk aversion—that was observed empirically by Paravisini et al. (2016) in peer-to-peer lending platforms. In the model, borrowers are jointly liable for the loan ℓ, as joint liability makes borrowers responsible for repaying each other's debt, which encourages risk sharing among the members who take a loan Attanasio et al. (2016). As Chen et al. (2017) highlight, this type of group lending lowers operating costs due to diligence and monitoring, and therefore increases the likelihood of loan repayment by shifting the bulk of monitoring costs from lenders to groups.

The performance of the business created by the non-defaulting members of the SFG and the counterfactual business of non-members is calculated using the stochastic business model below:

$$\begin{cases} \text{debt} = \ell(1 + i_\ell) & (4.28) \\ \text{inventory}_{(t=0)} = \ell(1 - \theta_\ell) & (4.29) \\ \text{sales}_t = m\,(\text{inventory}_t) & (4.30) \\ c_t = \text{sales}_t - q_t & (4.31) \\ \text{returns} = \sum_{t=1}^{\mathfrak{T}} c_t(1 - \epsilon_e) & (4.32) \\ \bar{a} = 2^{-1}\left(\ell + \text{inventory}_{(t=\mathfrak{T})} + \sum_{t=1}^{\mathfrak{T}} c_t\right) & (4.33) \\ ROA = (\bar{a}^{-1})\text{returns}. & (4.34) \end{cases}$$

In the model, the total debt capital of the business is the result of adding the amount of the loan borrowed by the SFG members plus the interest rate charged for the loan (equation 4.28). At the start of the business (at time $t = 0$), a fraction (θ_ℓ) of the borrowed loan is set aside to buy retail inventory (equation 4.29). The inventory reduction is a function of market sales (equation 4.30),

$$m = \left(1 + e^{\left(-1+e^{-\eta^2}\right)\left(-1+k_{s,(m-d)}\right)}\right)^{-\left(1+\frac{1}{\eta}\right)}, \tag{4.35}$$

which is boosted by the social capital of non-defaulting members ($k_{s,(m-d)}$) but can be lessened by the macroeconomic environment—the GDP growth—of a country (η).

Equation 4.31 is a mathematical description of a simple cash flow in the business: income is obtained by sales at time t, minus the loan repayment quota q_t. The returns at the end of the period $\mathfrak{T}$ (equation 4.32) are the sum of the cash flow minus random expenses related to unexpected events (ϵ_e). Returns on assets (ROA)—more precisely, returns on average assets—are obtained by dividing the business utility (equation 4.32) by the average assets of the business (equation 4.33). The formula of ROA in equation 4.34 is based on the business-success indicators suggested by the International Finance Corporation (2008) to evaluate micro, small and medium enterprises.

The simplified businesses model simulated in equations 4.28 to 4.34 is motivated by Herranz et al. (2015), who found that risk-averse entrepreneurs run smaller, more highly leveraged firms, which default less because running a smaller firm with higher debt reduces the number of personal funds at risk in the firm. In the model, a simplified balance sheet is assumed where assets are an addition of the income derived from sales plus the inventory and fixed assets acquired with the loan. The liabilities of the business are only the loan repayments q_t. Fixed assets are assumed to depreciate to zero at the end of the life-cycle of the business, and thus the utility at the end of the period is computed as the aggregate income from sales minus the total expenses incurred in paying the capital and interest of the loan, along with the expenses caused by unexpected (random) events.

Social capital enters the business model through improvements in market allocation pushed forward by the social capital of the borrowers in equation 4.30. Following Batjargal, 2003, the heterogeneity in the structural, relational, and resource-based aspects of social capital is reflected in various aspects of business performance because embedded relations influence the purchase and sale decisions of entrepreneurs. Also, as noted by Ling-Yee (2004), social capital helps to integrate the existing knowledge of members with the unique information from the market m. This in turn helps the group to update its knowledge, endow it with meaning, and translate it into organizational routines.

The counterfactual business simulation of non-members is also based on equations 4.28 to 4.34. The ABS-SFG model (randomly) chooses agents from the population of the village who were not part of the self-financing group. The selected agents create a business under the same financial conditions of the business created by the members of the self-financing group, i.e. the same loan amount and interest rate. Using the same financial conditions in both the self-financing business and the business financed with formal loans allows us to isolate the financial effects from the effects on business performance caused by social capital.

Box 4. Algorithm 4: Business simulation

Data: $k_m \ni \{k_\ell, k_s\}$, $\phi \equiv i$
Result: $\mathsf{ROA}_{\phi,s}$
for $\mathcal{S} \leftarrow s$ **do**
  $\ell_s = b_s(1 - r_{s,\text{post}})$
  $$m_s = \left(1 + e^{\left(-1+e^{-\eta^2}\right)\left(-1+k_{s,s,(m-d)}\right)}\right)^{-\left(1+\frac{1}{\eta}\right)}$$
  for $\Phi \leftarrow \phi$ **do**
    $i_{s,\ell} = \phi + r_{s,\text{post}}$
    $\text{debt}_s = \ell_s(1 + i_{s,\ell})$
    for $\mathfrak{T} \leftarrow t$ **do**
      $\text{inventory}_{s,(t=0)} = \ell_s(1 - \theta_\ell)$
      $\text{sales}_{s,t} = m_s\,(\text{inventory}_{s,t})$
      $c_{s,t} = \text{sales}_{s,t} - q_{s,t}$
      $\text{returns}_s = \sum_{t=1}^{\mathfrak{T}} c_{s,t}(1 - \epsilon_e)$
      $\bar{a}_s = 2^{-1}\left(\ell_s + \text{inventory}_{s,(t=\mathfrak{T})} + \sum_{t=1}^{\mathfrak{T}} c_{s,t}\right)$
    end
    $\mathsf{ROA}_{\phi,s} = (\bar{a}_s^{-1})\text{returns}_s$
  end
end

D.4 Results of computational experiments

This section runs $s = 1, 2, ..., \mathcal{S}$ simulations of the agent-based model of self-financing groups. The index s denotes running a single sequence of the whole model (the four algorithms described in Section D.3), and hence $\mathcal{S}$ is the total number of simulations of the ABS-SFG model. For example, when $s = \mathcal{S} = 1$, only a single village, one self-financing group, and one business are simulated—for an illustration of this simulation see the Appendix. If $\mathcal{S} = 1000$, then 1000 villages are randomly populated and 1000 different groups and businesses are generated in each village.

D.4.1 Counterfactual experiment of business performance

Tables 4.13 and 4.14 and Figure 4.11 show the results of a counterfactual experiment of business performance based on 1000 simulations of the ABS-SFG model ($s = 1, 2, ..., 1000; \mathcal{S} = 1 \times 10^3$). The experiment compares the returns on assets ($\mathsf{ROA}_{\phi,s}$) of 1000 businesses created by non-defaulting members of 1000 self-financing groups against 1000 businesses created by non-members in 1000 artificial communities.

The experiment simulates the impact of annual loan interest rates equal to 10% to 70% on $\mathsf{ROA}_{\phi,s}$ ($\phi \in \{.1, ..., \Phi = .7\}, \phi \equiv i$ in equation 4.27) for different values of savings contribution ρ in self-financing groups. The large values of the interest rates are based on the fact that borrowers in self-financing groups typically pay interest rates of 5% to 10% a month, according to

Rasmussen (2012). Table 4.11 shows the numerical values used to initialize the model. The main characteristics of the experiments are summarized in Table 4.12.

When annual interest rates are below 40%, the profitability of the businesses financed with loans from the self-financing groups is on average higher compared to the profitability of businesses financed with commercial loans (Table 4.13). When the savings quota is $\rho = 30\mathsf{MU}$, for example, the average return of the businesses in the self-financing groups is 5.89%, while the average return of the businesses financed with commercial loans is -3.31%. The risk of the businesses financed with commercial loans is also higher, equal to 5.67%, compared to the average risk of the businesses of the self-financing groups (4.25%).

For annual interest rates between 40% and 70% and for savings contributions of $\rho = 30$ and $\rho = 40$, negative returns are observed both for the businesses financed with commercial loans and for the business financed with commercial loans from self-financing groups. The businesses in the self-financing groups have positive returns only for quotas of savings equal to $\rho = 50$ and $\rho = 60$ (Table 4.14). For a savings quota of $\rho = 40\mathsf{MU}$ the average return of the businesses in the self-financing groups is -3.29%, while the average return of the businesses financed with commercial loans is -30.48%. If the savings quota rises to $\rho = 50\mathsf{MU}$, the returns of the businesses in the self-financing groups increases on average to 4.82%. This last result is the consequence of social capital in the self-financing group, which becomes important in the presence of costly debt capital.

Figure 4.11 reveals an emergent pattern in the dynamics of the returns of businesses financed by self-financing groups: bifurcation. For values of the savings quota ρ equal to $40\mathsf{MU}$ and $50\mathsf{MU}$, the businesses of the members of self-financing groups tend to outperform the profitability of the businesses of the non-members (Figures 4.11a and 4.11b). When the savings quota exceeds a threshold of $\rho = 60$ (Figure 4.11c), however, business performance splits into two branches (Figure 4.11d): in the lower bifurcation branch, the businesses of self-financing groups have average returns of -5.72% to 8.11%, while in the upper branch these businesses have returns of 20.32% to 27.47%. The risk, measured by the standard deviation of the returns, is also low in the lower branch of the simulated business of SFG members (Tables 4.13 and 4.14).

The bifurcation is caused by a quota of savings that exceeds a threshold of tolerance and creates nonlinear dynamics in the business profitability of self-financing groups. An extremely high quota of savings is a burden for agents with a restricted budget, which leads to savings default. Savings default is imitated due to the interaction of defaulting members with other group members, and as a consequence the group ends up having only a small fund available for loans. A lower amount of loans, in turn, leads to lower returns in the businesses created by the non-defaulting members, which generates the lower bifurcation branch in returns.

On the other hand, if members of the self-financing group manage to accommodate to the higher quota of savings and do not enter a default state, then the other members mimic their fulfilling behavior and hence at the end of the life-cycle of the group a larger fund is available for loans. The higher amount of loans, added to the social capital formed through homophily, boosts the profitability of the businesses created by self-financing groups, thereby generating an upper branch in the bifurcation pattern.

D.5 Conclusion

Atlan (1979, 1991) develops two complementary concepts of complexity: algorithmic complexity and contextual complexity. Algorithmic complexity is based on optimization, whereas contextual complexity is based on the communication among heterogeneous agents with conflicting goals Vasconcelos and Ramirez (2011). In this study, self-financing groups are considered a phenomenon of contextual complexity and an agent-based model is proposed to simulate how these groups form and create businesses in an artificial community.

The results allow us to conclude that the startup businesses of self-financing groups are more profitable and less risky compared to businesses financed with commercial loans, even with high interest rates, when social capital is properly consolidated. The consolidation of social capital is a consequence of the interaction among agents in the self-financing group. Social capital complements the debt capital in the fund available for loans, creating a competitive advantage that increases business profitability.

Higher quotas of savings in the group were found to boost profitability by raising the collective fund available for loans, but only up to a threshold, after which a bifurcation in returns appears. This bifurcation—typical in complexity dynamics see, Gao et al. (2016)—is a branching process of the dynamical system in which the topological structure switches to different states due to a change in a bifurcation parameter Crawford (1991). In the ABS-SFG model, the bifurcation parameter is the quota of savings agreed among members of a self-financing group. The bifurcation implies that field officers—hired by development agencies for the task of managing a group—face a trade-off between two possible states when raising the savings quota of a self-financing group: while the bifurcation parameter is a potential source of profit, increasing the quota of savings exacerbates also the risk of group failure.

The emerging findings of the study indicate that self-financing groups create a competitive advantage for business, as a consequence of the social capital formed in the group through homophily. Social capital, according to K.-H. Lin et al. (2016), constitutes an additional production factor that influences the competitive power and economic development of a venture, because social capital is based on network ties and thus it is a non-substitutable resource that cannot be acquired through imperfect imitation[19].

The theoretical implication of the findings is that homophily plays a dual role in self-financing groups. Following a resource-view approach, group formation is based on the maximization of utility by acquiring more resources in the form of loans and/or social capital. In this study, we argue that homophily plays a complementary role to utility maximization during the formation of a group. Homophily among members consolidates social bonds and reduces risk aversion during the life-cycle of a group. Social bonds translate into stronger cohesion, trust and peer pressure among members, which reduces the chances of default and facilitates organizational strategies[20].

The study also has managerial implications for traditional competencies, networking and market appreciation[21]. First, in a business of a self-financing group, traditional managerial competencies—such as finance, accounting, marketing, personnel management technologies, organizational procedures and routines Vasconcelos and Ramirez (2011)—are necessary to manage internal issues, and coordinate, motivate and select priorities. Second, due to the networking nature of self-financing groups, additional managerial competencies are required to construct value co-production systems on the basis of the collaboration and arrangements between members. Finally, through contextual listening, businesses of self-financing groups are able to appreciate, evaluate, question, and understand the general trends that compose the transactional environment.

Future studies can explore the business impact of self-financing groups that include a compo-

[19] The findings on the importance of social capital are consistent with the empirical study of Bosma et al. (2004), who find that investment in social capital enhances entrepreneurial performance of small businesses in terms of survival, profits, and generated employment. Torres et al. (2018) show also that social capital increases revenues and is a key asset for the long-term resilience of small businesses.

[20] While Bosma et al. (2004) relate the impact of social capital on firm performance to productivity and signaling, this study argues instead that the impact of social capital on the performance of businesses in self-financing groups is related to the cohesion created by homophily, which reduces organizational conflict. Previous studies found that intra-organizational social capital has a significant impact on the performance of new ventures Baum et al. (2000), because in business startups members are in unfamiliar roles and face new work relationships during a time of stretched resources. Nahapiet and Ghoshal (1998) further regard social capital as an organizational resource, and Stinchcombe (2000) propose that the performance of a new firm is significantly affected by the organizational conditions surrounding its founding. As Vasconcelos and Ramirez (2011) highlight, complexity, in this view, is a manageable dimension that can contribute to organizational learning.

[21] According to Vasconcelos and Ramirez (2011), management copes with complexity at three different levels: managerial competencies, networking and contextual listening.

nent of human capital besides social capital, as well as the role of friendship in business performance and the potential competitive advantages of a transactive memory system in self-financing groups[22]. While self-financing groups often focus on individual ventures, a joint business creates a competitive advantage for group members due to the combined effect of debt capital and the social capital generated through a dual process of homophily. Business training—which improves human capital—further encourages the competitive advantage of joint businesses.

Development agencies who work with self-financing groups as a platform to provide communities with sustainable development programs—like entrepreneurship, agriculture, adaptation to climate change, health and sanitation, or programs of literacy, education, and women empowerment—can use the ABS-SFG model as a cost-effective virtual laboratory to perform artificial experiments. The impact of intervention programs and social policies can be evaluated ex ante through the artificial experiments in the virtual laboratories. Investigations about the impact of business interventions are a promising research avenue, since Gonzales Martínez (2019) finds that business training is not the most frequent intervention offered to self-financing groups by development agencies, but is in fact the most important program to encourage financial sustainability, particularly after a development agency leaves the community where a group operates[23].

As shown in this study, agent-based modeling offers fascinating opportunities to understand and explore phenomena through a set of flexible computational tools. The simulations of agent-based models inform decision-making and allow one to formulate theories, that can guide empirical research and the interpretation of experimental evidence Chávez-Juárez (2017). In contrast to results estimated from observational data, the findings in agent-based models emerge from the interactions among heterogeneous agents in artificial worlds Gilbert and Troitzsch (2005); thus—as noted by Vermeulen and Pyka (2017) and Pyka et al. (2018)—agent-based modeling is a platform to experiment with complexity in a microverse of simulated realities.

[22]Self-financing groups can raise human and social capital through financial literacy. Engström and McKelvie (2017) argue that financial literacy addresses an individual's ability to internally assess the benefits and costs of an entrepreneurial opportunity. As Nguimkeu (2014) highlights, entrepreneurship requires not only financial capital but also human capital in the form of education, experience, and skills to develop ventures—see Radhakrishnan (2015).

Self-financing groups improve human capital during the meetings of the group by providing members with training in entrepreneurial skills and financial literacy. Engström and McKelvie (2017), after analyzing a dataset of 739 micro-enterprises in Ecuador, find that the impact of this training leads to improved financial performance of micro-enterprises in the informal economy. More recently Tsai, Yang, et al. (2018) found that human capital, measured by education and experience, improves vendor profit.

In the case of friendship, Batjargal (2003) finds that friendship ties affect firm performance negatively, because friendship leaves little room for maneuvering and creates financial concessions that harm a business's revenues and profit margins.

[23]Gonzales Martínez (2019) provides large-sample empirical evidence of the importance of business for self-financing groups, based on machine-learning methods. Theoretically, self-financing groups can improve business performance because these groups are a vehicle for the formation of a transactive memory system, which consists of the knowledge stored in each individual's memory combined with a metamemory containing information regarding the different teammates' domains of expertise. Xu (2016)—building on Wegner (1987), Uzzi (1997), Nadler et al. (2003), Borgatti and Cross (2003), and Argote et al. (2003)—indicates that strong relationships help the members of a group to develop transactive memory systems due to frequent interactions that facilitate reciprocal understandings of complex problems and consequently ease the transfer of complex information because of the norms of reciprocity and cooperation associated with social cohesion. As Xu (2016) concludes, this cognitive orientation influences how entrepreneurs develop a business plan, plan for a business operation, obtain funding to begin product/service development, and launch their startup.

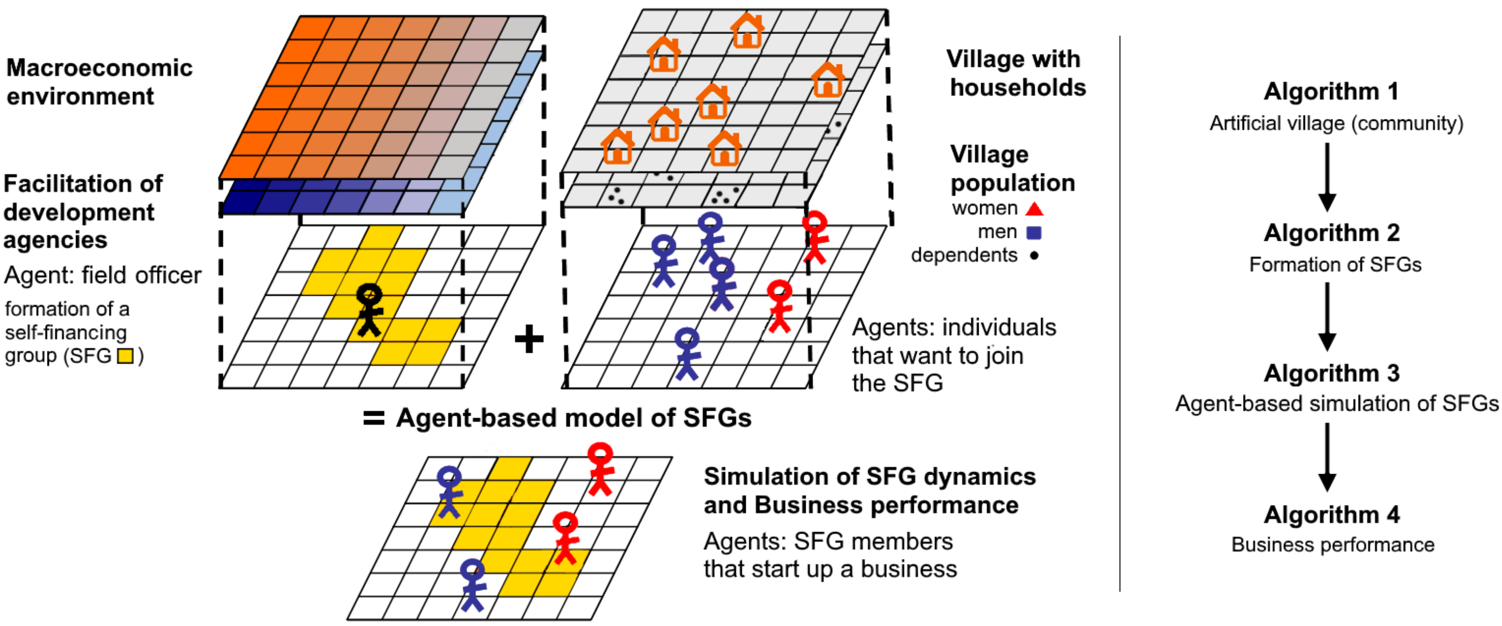


Figure 4.10: Schematic representation of the agent-based model of self-financing groups (adapted from Rebaudo et al., 2011). The figure illustrates the multilayered structure of the ABS-SFG model: SFG dynamics and business performance are the result of the interaction between external factors—the macroeconomic environment and the facilitation mechanisms of development agencies—plus the internal behavior of the individuals in the artificial community and the members of the self-financing group. The model is a sequence of four algorithms: Algorithm 1 randomly creates an artificial population inhabiting households in a village: working-age women, working-age men and household dependents (children and the elderly). Based on socioeconomic characteristics, homophily and intra-household conflict, some individuals of the working-age population want to join the SFG. In Algorithm 2, an agent hired by a development agency (the field officer) forms a SFG by choosing members from the individuals that want to be part of the SFG. Member selection is based on a gender rule (a preference for women). Algorithm 3 simulates the dynamics of the self-financing group: members allocate their savings into a common fund in each meeting and then take a joint loan from the accumulated fund. Social capital is created through homophily as the result of participating in the SFG. Algorithm 4 simulates the financial performance of a business started by the SFG members that do not fail to contribute with their savings. The performance of the business of the SFG is compared with the performance of a counterfactual business financed by a loan obtained from a formal financial institution.

Table 4.10: Structure of the agent-based simulation of self-financing businesses (ABS-SFG)

Phase/algorithm	Variables/traits in each submodel	Equations
Initialization		
Algorithm 1: Generation of an artificial community		
Households	○ Family size ○ Intra-household productive individuals ○ Intra-household gender composition	Centered Poisson distribution (eq. 1) Uniform discrete distribution (eq. 2) Uniform discrete distribution (eq. 3)
Village (community)	● Number of households in the village ○ Age profile in the village ○ Income profile in the village	None (initialization parameter) Mixture of uniform distributions (eq. 4) Log-normal distribution (eq. 5)
Algorithm 2: Formation of a self-financing group (SFG)		
Agents that want to join the SFG	○ Probability of joining a SFG ○ Social bonds (homophily) ○ Geographical distance (homophily) ○ Intra-household conflicts ○ Lack of access to financial services	Mixture of probabilities (eq. 6) Sigmoid function (eq. 7) Sigmoid function (eq. 8) Sigmoid function (eq. 9) Sigmoid function (eq. 10)
Field officer (agent)	○ Gender ratio of women in the SFG	Conditional function (eq. 11)
Running phase		
Algorithm 3: SFG dynamics of savings accumulation and formation of social capital		
Savings allocation across meetings	○ Probability of not contributing with savings ○ Group-level probability of default in savings ● Amount of savings quota of each member ● Mimicking behavior	Stochastic inequality (eq. 12) Hyperbolic tangent (eq. 13) None (simulation parameter) None (simulation parameter)
Members of the SFG (agents)	○ Idiosyncratic probability of default in savings ○ Income of SFG members ○ Age of SFG members ● Gender risk of not contributing to the SFG	Mixture of probabilities (eq. 14) Sigmoid function (eq. 15) Sigmoid function (eq. 16) None (simulation parameter)
Savings accumulation	○ Savings accumulation in the common box ○ Adaptative rule in the case of default ● Threshold of SFG failure	Accumulation of contributions (eq. 17) Stochastic addition of savings (eq. 18) None (simulation parameter)
Social capital	○ Social capital (function of homophily) ○ Impact of participating in SFG ○ Age differences among members ○ Income differences among members ○ Household distance among members	Homophily among members (eq. 19) Scale factor (eq. 20) Inverse Euclidean distance (eq. 21) Inverse Euclidean distance (eq. 22) Euclidean distance (eq. 23)
Algorithm 4: Loan provision and business simulation		
Loan allocation	○ Probability of receiving a loan ● Risk aversion (ex ante) ○ Risk aversion (ex post) ○ Amount of the loan allocated to borrowers ○ Effective interest rate charged to loans	Social capital and income (eq. 24) None (simulation parameter) Updated risk aversion (eq. 25) Fraction of total savings in the box (eq. 26) Interest rate plus risk aversion (eq. 27)
Business performance	● Fraction of assets allocated to inventory ● Impact of the macroeconomic environment ○ Total amount of debt (principal + interest) ○ Inventory ○ Retail sales ○ Cash flow ○ Utility (returns) ○ Average assets ○ Returns on assets (ROA) ○ Market sales	None (simulation parameter) None (simulation parameter) Debt function (eq. 28) Initial inventory (eq. 29) Income gained from sales (eq. 30) Income flow minus loan repayments (eq. 31) Returns minus random expenses (eq. 32) Assets over the period (eq. 33) Returns divided by average assets (eq. 34) Function of social capital (eq. 35)

Notes on simulation values:
(●) Predetermined
(○) Produced by the model

Table 4.11: Numerical values used to initialize the ABS-SFG model

Phase/algorithm	Numerical values	Notes on parameter values
Initialization		
Algorithm 1: Generation of an artificial community		
Family size	$\lambda_h = 2$	Parameter of a centered Poisson distribution.
Intra-household productive individuals and gender composition	$u_\delta = 3$	Upper parameter of a uniform integer discrete distribution. The lower parameter is always one because there is always one productive individual in a productive household
Number of households in the village	64	Number of households in a village
Age profile in the village	$u_{age} = 20$	Upper parameters of a mixture of discrete uniform distributions
Income profile in the village	$\mu_{\neg w} = 5.5$ $\mu_w = 5.8$ $\sigma_{\neg w} = 0.5$ $\sigma_w = 0.4$	First and second parameter of a log-normal distribution. Lower values of μ (compared to those of men) imply that the central tendency of the income distribution of women is lower than that of men. Higher values of σ imply that the income differences are more dispersed across individuals
Algorithm 2: Formation of a self-financing group (SFG)		
Overall probability of joining a SFG	$\omega_{\mathbb{P}_i(m)} = .55$	Weight (importance) of lack of access to financial services for the agents that want to be part of the SFG
Gender ratio of women in the SFG	$\tau_w = 0.7$	Minimum percentage of women in a group required by the field officer
Running phase		
Algorithm 3: SFG dynamics of savings accumulation and formation of social capital		
Mimicking behavior	$\beta = .007$	Larger values increase the probability that non-defaulting members will enter a default state when a member fails to deliver her/his quota of savings
Idiosyncratic probability of default in savings	$\omega_{d_s} = 0.5$	Weight (importance) of income and age in the probability of not contributing with savings
Gender risk of not contributing to the SFG	$\gamma_g = 0.7$	Women have less probability of failing to contribute their savings, compared to men
Threshold of SFG failure	$\tau_d = 0.2$	Maximum tolerance for the percentage of members failing to contribute their savings
Algorithm 4: Loan provision and business simulation		
Risk aversion (ex ante)	$\omega_r = 0.8$	Initial (ex ante) risk aversion of the group against allocating loans. This risk aversion is updated after the members experience being part of a SFG
Fraction of assets allocated to inventory	$\theta_\ell = 0.65$	Larger values imply that a higher proportion of the loan amount will be used to buy inventory for retail sales
Impact of the macroeconomic environment	$\eta = 0.05$	Impact of economic growth on market sales. Larger (smaller) values will increase (decrease) the business sales

Table 4.12: Main characteristics of the experiments

Parameter	Values
Nominal interest rate ($\phi \equiv i$)	10% to 70%
Savings quota (ρ)	30, 40, 50, 60, 70

Table 4.13: Business simulation results: Loan interest rate in the range of 10% to 39.9%

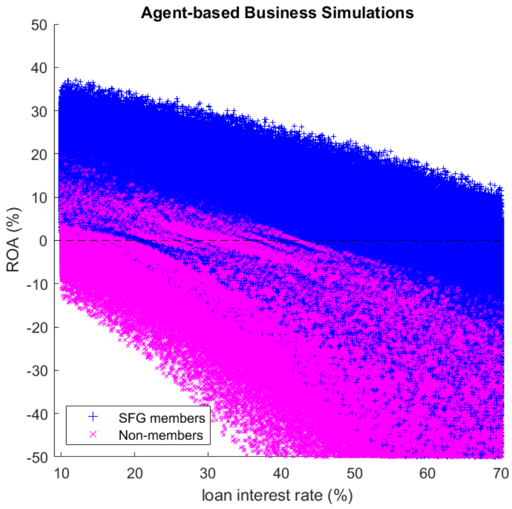
Savings quota	Business impact	ROA (%)	
		Members	Non-members
$\rho = 30$	Average returns	5.89	-3.31
	Risk	4.25	5.67
$\rho = 40$	Average returns	11.95	-3.17
	Risk	3.43	5.65
$\rho = 50$	Average returns	17.03	-0.18
	Risk	2.83	5.24
$\rho = 60$	Average returns	25.21	18.03
	Risk	2.10	2.92
$\rho = 70$	Average returns	[27.47, 8.11]	21.35
	Risk	(2.16, 2.82)	2.61

Note: When $\rho > 60$, a bifurcation pattern appears in returns
ROA: returns on assets
Average returns: average ROA in the 1000 simulations
Risk: standard deviation of ROA in the 1000 simulations

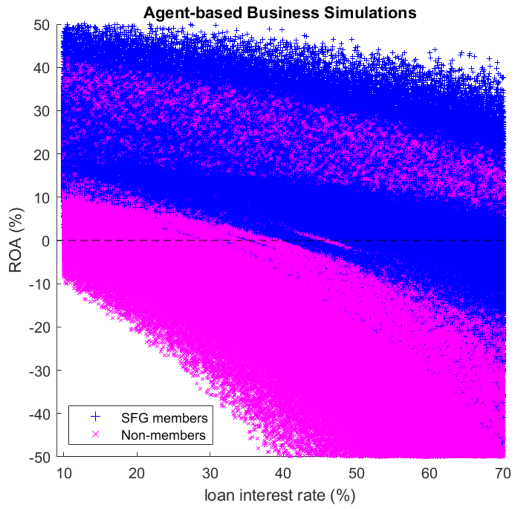
Table 4.14: Business simulation results: Loan interest rate in the range of 40% to 70%

Savings quota	Business impact	ROA (%)	
		Members	Non-members
$\rho = 30$	Average returns	-13.58	-30.76
	Risk	7.44	11.01
$\rho = 40$	Average returns	-3.29	-30.48
	Risk	5.65	10.95
$\rho = 50$	Average returns	4.82	-25.29
	Risk	4.42	10.03
$\rho = 60$	Average returns	16.66	5.25
	Risk	3.03	4.75
$\rho = 70$	Average returns	[20.32, -5.72]	10.84
	Risk	(2.67, 6.22)	3.90

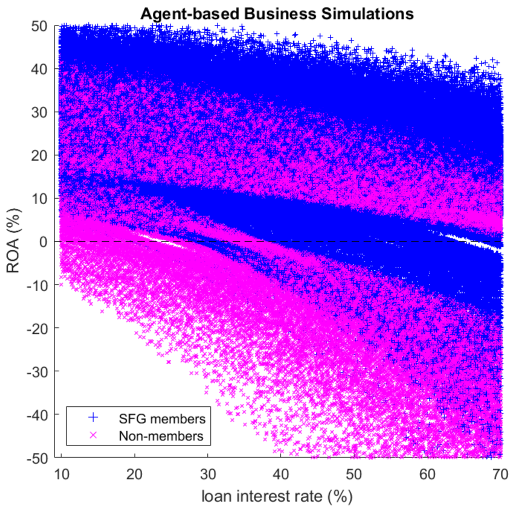
Note: When $\rho > 60$, a bifurcation pattern appears in returns
ROA: returns on assets
Average returns: average ROA in the 10000 simulations
Risk: standard deviation of ROA in the 10000 simulations



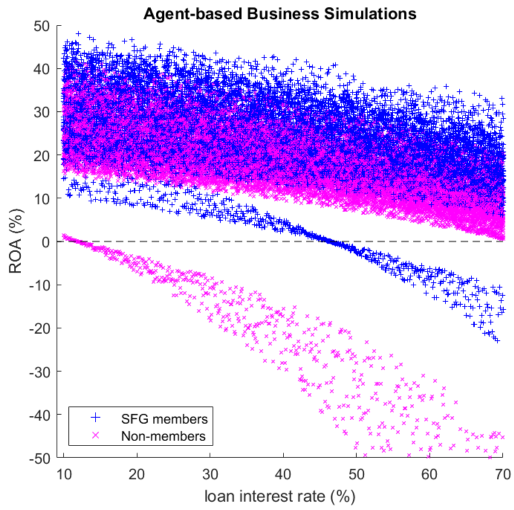
(a) Savings quota $\rho = 40$



(b) Savings quota $\rho = 50$



(c) Savings quota $\rho = 60$



(d) Savings quota $\rho = 70$

Figure 4.11: Agent-based simulation of business performance. When $\rho < 60$, the performance of the businesses of the members of self-financing groups is higher compared to the performance of the businesses financed with commercial loans. After the bifurcation point ($\rho > 60$), in the lower bifurcation branch groups fail to generate enough financial capital—because many members fail to contribute to the common fund—and their businesses perform worse than those of non-members. In the upper bifurcation branch, members adapt to the high quota of savings and create additional social capital, boosting the performance of their business initiatives.

Appendix: Illustration of the dynamics of the ABS-SFG model with one single simulation

This appendix illustrates the dynamics of the ABS-SFG model by showing the results of running only one simulation. Figure A1 shows the results of simulating one artificial community with $H = 64$ households. Each box represents a household. Blue squares in the households are productive men, red triangles are women, and black dots are the dependents in the household (children and non-working elderly populations). The parameter H calibrates the number of households in a community.

In some households there is only one woman or man and one dependent agent, while in other households there is more than one agent of the working population and also several dependents (Figure A1). In the simulated community, there is a total of 364 individuals, 175 of which are men and 189 are women (the gender ratio is .93). In the village, 232 individuals are dependents agents and 132 are agents of the labor force.

Figure A1 (left) also shows the distributions of age and income in the artificial village. The distribution of income is skewed—a common feature of income distributions—with a bulk of individuals in the average income and some individuals with high income in the right tail of the distribution. The income distribution of men is set higher compared to that of women, in order to simulate gender disparities in income commonly found in empirical studies. Figure A1 (left) shows also that the average age in the population is 43 years, with some individuals having less than 20 years and others having close to 70 years. The age dependency ratio is 1.76 in the simulated community, reflecting the fact that the community has more dependents than workers.

The number of members of a self-financing group is an emergent parameter of the agent-based model. In a single simulation for $\rho = 40$ with a fixed seed (Figure A2), a group of 18 members is created by an artificial field agent. In the group, 15 agents are women and 3 are men, indicating the preference of the field agent for women. The members of the group selected by the field agent yield a gender composition of the self-financing group equal to $\tau = 15/18 \approx 83\%$ of women.

In terms of group dynamics, Figure A2b shows that in the fourth meeting a member of the group (a man) fails to contribute his quota of savings. Due to the mimicking behavior of agents, other members of the self-financing group also start to fail to contribute to the common fund by the middle of the life-cycle of the group (Figure A2c). By the end of the life-cycle of the group—in the last meeting—only 9 members—7 women and 2 men—have not failed to contribute to the common fund of the self-financing group (Figure A2d).

Table A1 and Figure A3 show the impact of changing the amount of savings quota ρ that each member has to contribute to the self-financing group. When the members of the group contribute an individual quota of 30 monetary units (MUs), $\rho = 30MU$, there is no default since no member fails to contribute to the fund (Figure A3a). For a quota of $\rho = 40MU$, half of the members in the group fail to contribute to the fund (Figure A3b). When the quota is $\rho = 50MU$, 15 members fail to contribute to the fund (Figure A3c) and when the quota is $\rho = 60MU$ the individual savings contribution is too high and all of the members fail to contribute, leading to the failure of the group (Figure A3d).

The simulations of the impact of ρ show that higher quotas of savings can increase the common loan fund, up to a point beyond which raising the quota starts to reduce the common fund. An extremely high quota of savings causes members to default, which eventually decreases the fund available for startup loans. The number of defaulting members is related both to the individual circumstances of each agent and also to the interaction among agents. At the individual level, an extremely high quota of savings creates a heavy burden for the members of the self-financing group, due to household budget constraints. At the group level, due to a mimicking behavior and the stochastic interaction among agents Kirman (2010), agents have fewer incentives to contribute to the common fund if they observe that other agents are failing to contribute to the fund.

Finally, Figure A4 shows the social bonds of the agents before and after joining the self-financing group, for a quota of $\rho = 40MU$. Homophily—due to age, income, and household location—generates links among agents that increase the social capital in the group. The growth of social capital in the self-financing group is caused by the fact that during the life-cycle of the group, agents repeatedly meet with each other, strengthening their bonds. This is particularly true for those members that already had a tight social network before joining the group; for example, the female agents 16 and 15, and to a lesser extent the male member 7; see Figure A4). Figure A4 also reveals the inequality in social capital that the model aims to capture; i.e., agents have heterogeneous levels of social capital depending on their socioeconomic and geographical characteristics, as noted by Hsung et al. (2017).

Table A1: Results of the agent-based simulation on group dynamics

	Savings quota			
	$\rho = 30$	$\rho = 40$	$\rho = 50$	$\rho = 60$
Total amount collected in the fund (MU)	9720	8520	5784	4961
Number of default members	0	9	15	18
Non-default members (end of the cycle)	18	9	3	0

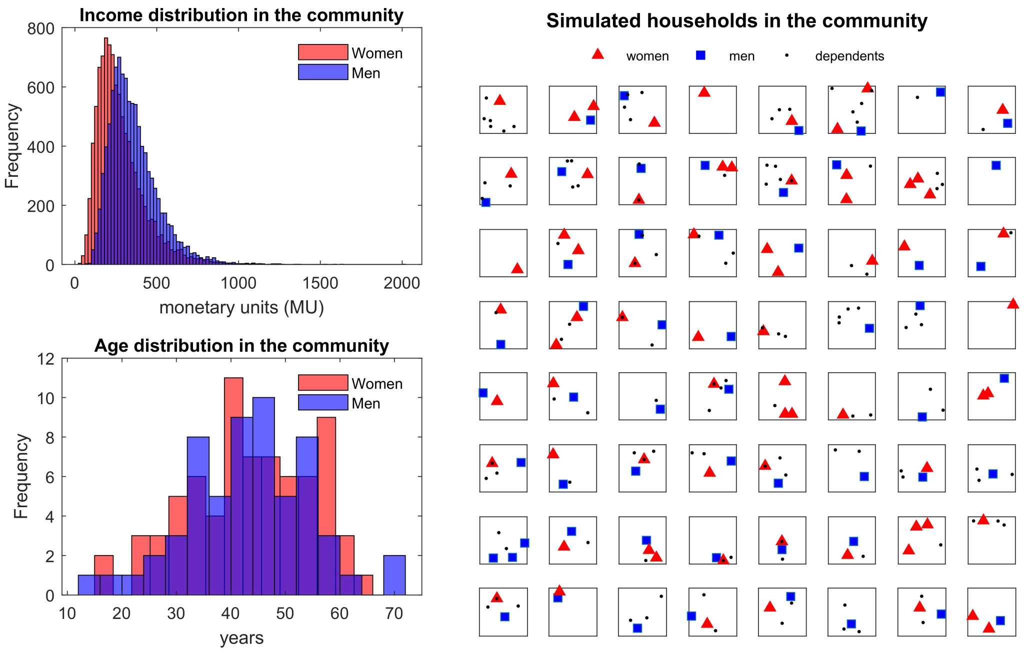
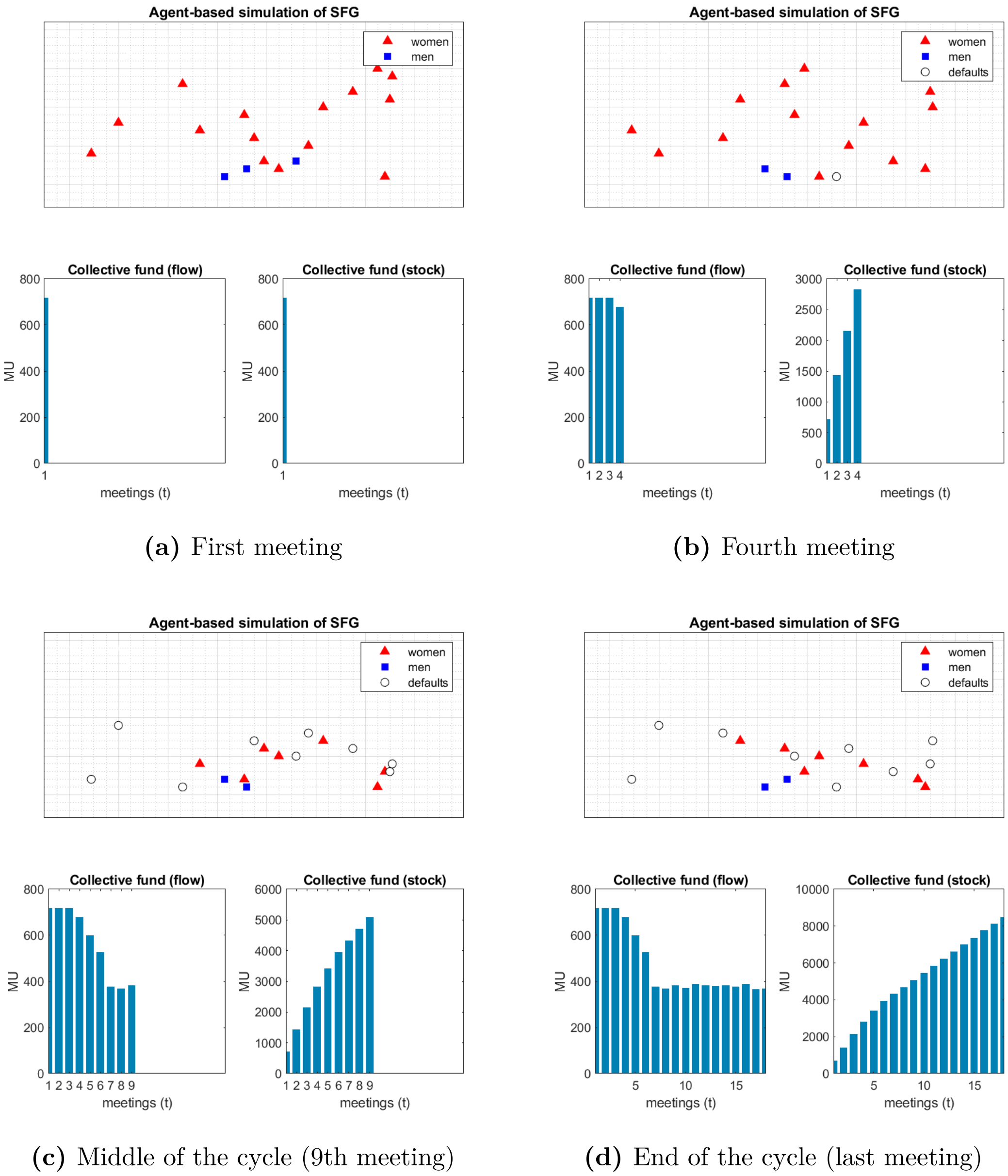


Figure A1: Simulation of a community of H = 64 households. Each box represents a household. The households are inhabited by men (blue squares) women (red triangles) and children and non-working elderly populations (black dots). H can be modified to simulate smaller villages with few households or larger villages. Due to the random creation of villages, different simulations produce different household compositions. The high proportion of low-income individuals in the village and the gender disparities related to a higher income inequality for women can be seen in the left-skewed distributions of income in the village (Figure A1 left). The distribution of age indicates a concentration of the population between 30 and 50 years, but with a high number of dependents compared to the labor force population, as shown in the demographic indicators below:
Population: 364 individuals
Men in the community: 175
Women in the community: 189
Gender ratio: 0.93
Dependents in the community: 232
Labor force (productive population): 132
Income distribution of the agents in the community.
Men income (median): 328.67 mu
Women income (median): 243.42 mu
Age distribution of the agents in the community.
Average age of the productive pop: 43
Age dependency ratio: 1.76



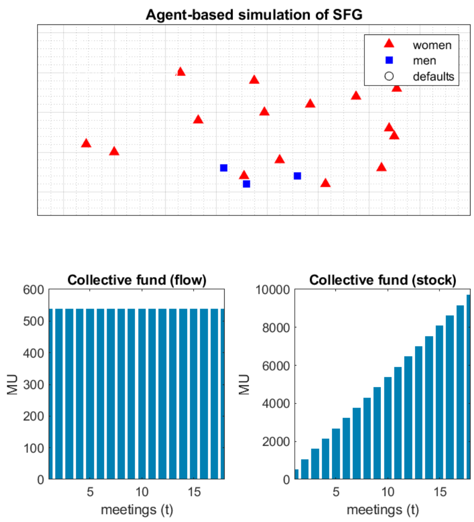
(a) First meeting

(b) Fourth meeting

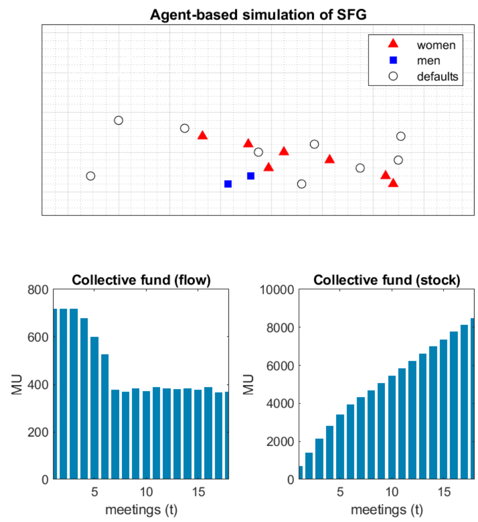
(c) Middle of the cycle (9th meeting)

(d) End of the cycle (last meeting)

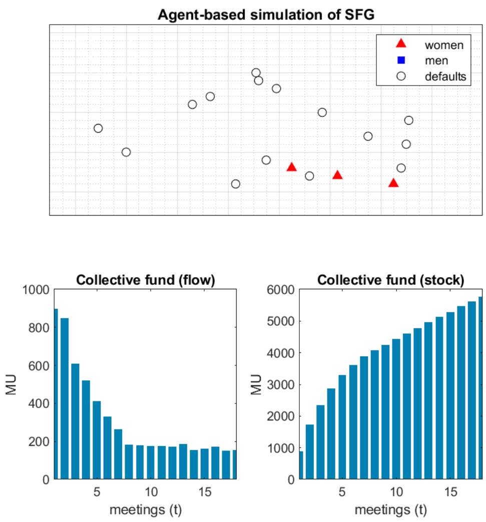
Figure A2: Agent-based simulation: one single self-financed group, savings quota $\rho = 40$. All the members contribute their savings to the common fund in the first meeting of the group. A male agent fails to contribute from his savings in the fourth meeting, and due to the mimicking behavior of other agents, 9 of the 18 original members of the group end up failing to contribute their savings to the common fund of the self-financing group.



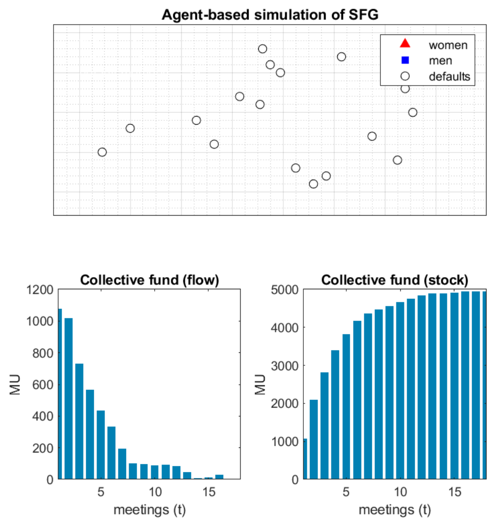
(a) Savings quota $\rho = 30$



(b) Savings quota $\rho = 40$



(c) Savings quota $\rho = 50$



(d) Savings quota $\rho = 60$

Figure A3: Agent-based simulation of a self-financing group with different values of savings quota (ρ). Low values of the individual savings contribution ($\rho = 30$) are not a burden for members of self-financing groups, but when the savings contribution increases to $\rho = 60$, members start to fail to contribute to the common fund, due to household budget constraints and the mimicking behavior of agents.

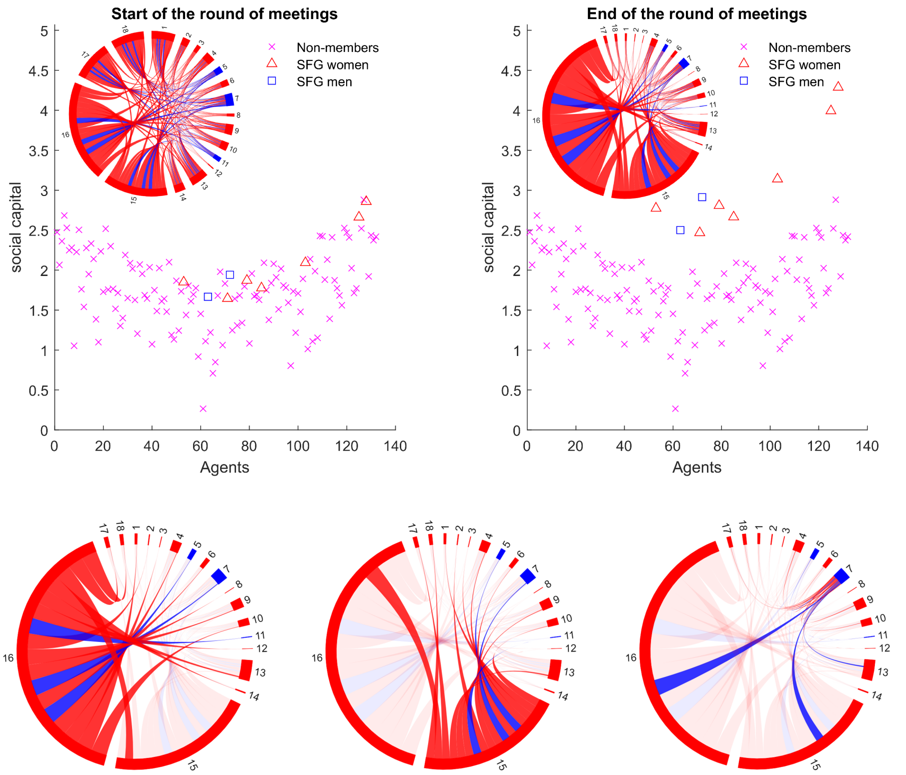


Figure A4: Social capital in a self-financing group with a savings quota of $\rho = 40\mathsf{MU}$. At the start of the life-cycle of the group, non-members and members of the self-financing group have similar social capital. At the end of the life-cycle of the group, non-defaulting members have higher social capital due to the strengthening of social bonds during the meetings. Female agents 16 and 15 have stronger social bonds with other members compared to other female and male agents (as agent 7). Thus, the heterogeneity in social bonds leads to inequality in social capital.

References

Abbena, E., Salamon, S., & Gray, A. (2017). *Modern differential geometry of curves and surfaces with mathematica*. Chapman; Hall/CRC.

Abbink, K., Irlenbusch, B., & Renner, E. (2006). Group size and social ties in microfinance institutions. *Economic Inquiry*, *44*(4), 614–628.

Acs, Z. J., & Armington, C. (2006). *Entrepreneurship, geography, and american economic growth*. Cambridge University Press.

Alila, P. O. (1998). Informal credit and rural small enterprises growth. *Journal of Asian and African studies*, *33*(2), 159–180.

Allen, H. (2006). Village savings and loans associations — sustainable and cost-effective rural finance. *Small Enterprise Development*, *17*, 61–68.

Allen, H., & Panetta, D. (2010). Savings groups: What are they. *Washington DC: SEEP Network*, 2.

Anderson, S., & Baland, J.-M. (2002). The economics of roscas and intrahousehold resource allocation. *The Quarterly Journal of Economics*, *117*(3), 963–995.

Annan, J., Bundervoet, T., Seban, J., & Costigan, J. (2013). *A randomized impact evaluation of village savings and loans associations and family-based interventions in burundi* (tech. rep.). USAID - International Rescue Committee.

Argote, L., McEvily, B., & Reagans, R. (2003). Managing knowledge in organizations: An integrative framework and review of emerging themes. *Management science*, *49*(4), 571–582.

Atlan, H. (1979). Entre le cristal et la fumée. *Essai sur l'organisation du vivant*.

Atlan, H. (1991). L'intuition du complexe et ses théorisations. *Ees Théories de la complexité. Autour de l'œuvre d'Henri Atlan. Paris, Seuil*, 9–42.

Atmadja, A. S., Sharma, P., & Su, J.-J. (2018). Microfinance and microenterprise performance in indonesia: An extended and updated survey. *International Journal of Social Economics*, *45*(6), 957–972.

Attanasio, O., Augsburg, B., & De Haas, R. (2016). Microcredit contracts, risk diversification and loan take-up. *Journal of the European Economic Association*.

Ban, R., Gilligan, M. J., & Rieger, M. (2015). *Self-help groups, savings and social capital: Evidence from a field experiment in cambodia* (World Bank Policy Research Working Paper No. 7382). The World Bank.

Banerjee, A., Yakovenko, V. M., & Di Matteo, T. (2006). A study of the personal income distribution in australia. *Physica A: statistical mechanics and its applications*, *370*(1), 54–59.

Batjargal, B. (2003). Social capital and entrepreneurial performance in russia: A longitudinal study. *Organization Studies*, *24*(4), 535–556.

Baudrillard, J. (1994). *Simulacra and simulation*. University of Michigan press.

Baum, J. A., Calabrese, T., & Silverman, B. S. (2000). Don't go it alone: Alliance network composition and startups' performance in canadian biotechnology. *Strategic management journal*, *21*(3), 267–294.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for sme finance. *Journal of Banking & Finance*, *30*(11), 2945–2966.

Biekpe, N. (2004). Financing small businesses in sub-saharan africa. *Journal of African Business*, *5*(1), 29–44.

Biggart, N. W. (2001). Banking on each other: The situational logic of rotating savings and credit associations. *Advances in Qualitative Organization Research*, *3*(1), 129–52.

Bisrat, A., Kostas, K., & Feng, L. (2012). Are there financial benefits to join roscas? empirical evidence from equb in ethiopia. *Procedia Economics and Finance*, *1*, 229–238.

Borgatti, S. P., & Cross, R. (2003). A relational view of information seeking and learning in social networks. *Management science*, *49*(4), 432–445.

Bosma, N., van Praag, M., Thurik, R., & de Wit, G. (2004). The value of human and social capital investments for the business performance of startups. *Small Business Economics*, *23*(3), 227–236.

Bouman, F. (1995). Rotating and accumulating savings and credit associations: A development perspective. *World Development*, *23*.

Bourdieu, P. (1986). The forms of capital, In *Handbook of theory and research for the sociology of education*. Greenwood Press.

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology*. Chicago und London: The University of Chicago Press.

Brody, C. M., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. (2015). Economic self-help group programs for improving women's empowerment: A systematic review. *Touro University California*.

Burlando, A., & Canidio, A. (2017). Does group inclusion hurt financial inclusion? evidence from ultra-poor members of ugandan savings groups. *Journal of Development Economics*, *128*, 24–48.

Butler, T., & Robson, G. (2001). Social capital, gentrification and neighbourhood change in london: A comparison of three south london neighbourhoods. *Urban studies*, *38*(12), 2145–2162.

Chávez-Juárez, F. (2017). On the role of agent-based modeling in the theory of development economics. *Review of Development Economics*, *21*(3), 713–730.

Chen, J., Chang, A. Y., & Bruton, G. D. (2017). Microfinance: Where are we today and where should the research go in the future? *International Small Business Journal*, *35*(7), 793–802.

Cheng, S. (2015). Potential lending discrimination? insights from small business financing and new venture survival. *Journal of Small Business Management*, *53*(4), 905–923.

Collins, A. J., & Frydenlund, E. (2018). Strategic group formation in agent-based simulation. *Simulation*, *94*(3), 179–193.

Cooke, P., & Wills, D. (1999). Small firms, social capital and the enhancement of business performance through innovation programmes. *Small Business Economics*, *13*(3), 219–234.

Crawford, J. D. (1991). Introduction to bifurcation theory. *Review of Modern Physics*, *63*, 991–1037.

Demirgüç-Kunt, A., Beck, T., & Honohan, P. (2008). Access to finance and development: Theory and measurement. *Finance for All? Policies and Pitfalls in Expanding Access*, 21–54.

D'espallier, B., Guérin, I., & Mersland, R. (2011). Women and repayment in microfinance: A global analysis. *World development*, *39*(5), 758–772.

Edwards, B., & Foley, M. (1997). *Social capital, civil society, and contemporary democracy*. Sage publications.

El-Sayed, A. M., Scarborough, P., Seemann, L., & Galea, S. (2012). Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, *9*(1), 1.

Engström, P., & McKelvie, A. (2017). Financial literacy, role models, and micro-enterprise performance in the informal economy. *International Small Business Journal*, *35*(7), 855–875.

Evelyn, U., & Osifo, S. J. (2018). Microfinance and female entrepreneurship in nigeria. *DBA Africa Management Review*, *8*(1).

Farmer, J. D., Patelli, P., & Zovko, I. I. (2005). The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences*, *102*(6), 2254–2259.

Feigenberg, B., Field, E., Pande, R., Rigol, N., & Sarkar, S. (2014). Do group dynamics influence social capital and female empowerment? experimental evidence from microfinance. *Journal of Policy Analysis and Management*, *33*(4), 932–949.

Field, E., Pande, R., Papp, J., & Rigol, N. (2013). Does the classic microfinance model discourage entrepreneurship among the poor? experimental evidence from india. *American Economic Review*, *103*(6), 2196–2226.

Flynn, J., & Sumberg, J. (2018). Are savings groups a livelihoods game changer for young people in africa? *Development in Practice*, *28*(1), 51–64.

Forrest, R., & Kearns, A. (2001). Social cohesion, social capital and the neighbourhood. *Urban Studies*, *38*(12), 2125–2143.

Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*. Free Press Paperbacks.

Gao, J., Barzel, B., & Barabási, A.-L. (2016). Universal resilience patterns in complex networks. *Nature*, *530*(7590), 307.

Garikipati, S. (2008). The impact of lending to women on household vulnerability and women's empowerment: Evidence from india. *World Development*, *36*(12), 2620–2642.

Ghosh, S., & Vinod, D. (2017). What constrains financial inclusion for women? evidence from indian micro data. *World Development*, *92*, 60–81.

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK).

Gonzales Martínez, R. (2019). Which social program supports sustainable grassroot finance? machine-learning evidence. *International Journal of Sustainable Development & World Ecology*, 1–7.

Gonzales Martınez, R., Aguilera-Lizarazu, G., Rojas-Hosse, A., & Blanco, P. A. (2019). The interaction effect of gender and ethnicity in loan approval: A bayesian estimation with data from a laboratory field experiment. *Review of Development Economics*.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, *91*(3), 481–510.

Greaney, B. P., Kaboski, J. P., & Van Leemput, E. (2016). Can self-help groups really be self-help? *The Review of Economic Studies*, *83*(4), 1614–1644.

Guha, S., & Gupta, G. (2005). Microcredit for income generation: The role of rosca. *Economic and political weekly*, 1470–1473.

Gurov, S. (2005). Reliability estimation of classification algorithms. ii. point bayesian estimates. *Computational Mathematics and Modeling*, *16*(2), 169–178.

Guterman, D., Harmon, D., & Roiland, J. (2015). The ricks must be crazy [S(2), 6].

Hakansson, H., & Snehota, I. (1995). Developing relationships in business networks. *International Thomson Business Press, London*.

Hakim, C. (2016). *Key issues in women's work: Female diversity and the polarisation of women's employment*. Routledge-Cavendish.

Haltiwanger, J., Jarmin, R. S., Kulick, R., Miranda, J., Et al. (2016). High growth young firms: Contribution to job, output, and productivity growth. *NBER Chapters*, 11–62.

Hamill, L., & Gilbert, N. (2016). *Agent-based modelling in economics* (1st). Wiley Publishing.

Herranz, N., Krasa, S., & Villamil, A. P. (2015). Entrepreneurs, risk aversion, and dynamic firms. *Journal of Political Economy*, *123*(5), 1133–1176.

Hospes, O. (1995). Women's differential use of roscas in indonesia., In *Money-go-rounds : The importance of rotating savings and credit associations for women*.

Hsung, R.-M., Fu, Y.-c., & Lin, N. (2017). The position generator: Measurement techniques for investigations of social capital, In *Social capital*. Routledge.

International Finance Corporation. (2008). *Financing micro, small, and medium enterprises: An independent evaluation of ifc's experience with financial intermediaries in frontier countries*. International Finance Corporation, World Bank Group.

Jennings, V., & Lloyd-Smith, B. (2015). Poisson and household size. *Mathematical Scientist*, *40*(2), 103–117.

Jennings, V., Lloyd-Smith, B., & Ironmonger, D. (1999). Household size and the poisson distribution. *Journal of the Australian Population Association*, *16*(1-2), 65.

Jiang, J. Y. a., & Liu, C.-W. (2015). High performance work systems and organizational effectiveness: The mediating role of social capital. *Human Resource Management Review*, *25*(1), 126–137.

Kast, F., Meier, S., & Pomeranz, D. (2012). *Under-savers anonymous: Evidence on self-help groups and peer pressure as a savings commitment device* (tech. rep.). National Bureau of Economic Research.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. OUP USA.

Khambule, I., & Siswana, B. (2017). How inequalities undermine social cohesion: A case study of south africa.

Kirman, A. (2010). *Complex economics: Individual and collective rationality*. Routledge.

Ksoll, C., Lilleør, H. B., Lønborg, J. H., & Rasmussen, O. D. (2016). Impact of village savings and loan associations: Evidence from a cluster randomized trial. *Journal of Development Economics*, *120*, 70–85.

Laird, A. K. (1964). Dynamics of tumour growth. *British journal of cancer*, *18*(3), 490.

Langton, C. G. (1997). *Artificial life: An overview*. Mit Press.

Lauder, W., Mummery, K., & Sharkey, S. (2006). Social capital, age and religiosity in people who are lonely. *Journal of Clinical Nursing*, *15*(3), 334–340.

le Polain, M., Sterck, O., & Nyssens, M. (2018). Interest rates in savings groups: Thrift or threat? *World Development*, *101*, 162–172.

Lemos, C. M. (2017). *Agent-based modeling of social conflict: From mechanisms to complex behavior* (1st). Springer Publishing Company, Incorporated.

Lin, K.-H., Ho, S.-H., & Peng, Y.-P. (2016). The application of social capital to the construction of organizational capability, In *Asian businesses in a turbulent environment*. Springer.

Lin, N. (2001). *Social capital: A theory of social structure and action* (Vol. 19). Cambridge university press.

Ling-Yee, L. (2004). An examination of the foreign market knowledge of exporting firms based in the people's republic of china: Its determinants and effect on export intensity. *Industrial Marketing Management*, *33*(7), 561–572.

Loury, G. Et al. (1977). A dynamic theory of racial income differences. *Women, minorities, and employment discrimination*, *153*, 86–153.

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, *4*(3), 151–162.

Mackenbach, J. D., Lakerveld, J., van Lenthe, F. J., Kawachi, I., McKee, M., Rutter, H., Glonti, K., Compernolle, S., De Bourdeaudhuij, I., Feuillet, T., Et al. (2016). Neighbourhood social capital: Measurement issues and associations with health outcomes. *Obesity reviews*, *17*, 96–107.

Mayoux, L. Et al. (2000). *Micro-finance and the empowerment of women: A review of the key issues* (tech. rep.). International Labour Organization.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Nadler, J., Thompson, L., & Boven, L. V. (2003). Learning negotiation skills: Four models of knowledge creation and transfer. *Management Science*, *49*(4), 529–540.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of management review*, *23*(2), 242–266.

Newman, A., Schwarz, S., & Ahlstrom, D. (2017). Microfinance and entrepreneurship: An introduction. *International Small Business Journal*, *35*(7), 787–792.

Nguimkeu, P. (2014). A structural econometric analysis of the informal sector heterogeneity. *Journal of Development Economics*, *107*, 175–191.

OECD. (2014). Policy brief on access to business start-up finance for inclusive entrepreneurship: Entrepreneurial activities in europe. Luxembourg: Publications Office of the European Union.

Order, H. (1995). How adaptation builds complexity. *Reading, MA: Perseus*.

Paravisini, D., Rappoport, V., & Ravina, E. (2016). Risk aversion and wealth: Evidence from person-to-person lending portfolios. *Management Science*, *63*(2), 279–297.

Pillai, D., & Abraham, S. (2017). Mediating role of self help groups for stimulating rural financial intermediation in india. *The Journal of Internet Banking and Commerce*, 1–11.

Ponta, L., & Cincotti, S. (2018). Traders' networks of interactions and structural properties of financial markets: An agent-based approach. *Complexity*, *2018*.

Putnam, R. (1993). The prosperous community: Social capital and public life. *The american prospect*, *13*.

Pyka, A., Mueller, M., & Kudic, M. (2018). Regional innovation systems in policy laboratories. *Journal of Open Innovation: Technology, Market, and Complexity*, *4*(4), 44.

Radhakrishnan, S. (2015). Low profile or entrepreneurial? gender, class, and cultural adaptation in the global microfinance industry. *World Development*, *74*, 264–274.

Rasmussen, O. D. (2012). Small groups, large profits: Calculating interest rates in community-managed microfinance. *Enterprise Development and Microfinance*, *23*(4), 298–318.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological methodology*, *29*(1), 1–41.

Rebaudo, F., Crespo-Pérez, V., Silvain, J.-F., & Dangles, O. (2011). Agent-based modeling of human-induced spread of invasive species in agricultural landscapes: Insights from the potato moth in ecuador. *Journal of Artificial Societies and Social Simulation*, *14*(3), 7.

Santana, L. (2017). Determinantes de la supervivencia de microempresas en bogotá: Un análisis con modelos de duración. *Innovar*, *27*(64), 51–62.

Sengupta, R. R., & Bennett, D. A. (2003). Agent-based modelling environment for spatial decision support. *International Journal of Geographical Information Science*, *17*(2), 157–180.

Shahriar, A. Z. M., Schwarz, S., & Newman, A. (2016). Profit orientation of microfinance institutions and provision of financial capital to business start-ups. *International Small Business Journal*, *34*(4), 532–552.

Smallbone, D., Ram, M., Deakins, D., & Aldock, R. B. (2003). Access to finance by ethnic minority businesses in the uk. *International Small Business Journal*, *21*(3), 291–314.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104.

Souma, W. (2001). Universal structure of the personal income distribution. *Fractals*, *9*(04), 463–470.

Spring, A. (2009). African women in the entrepreneurial landscape: Reconsidering the formal and informal sectors. *Journal of African Business*, *10*(1), 11–30.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American economic review*, *71*(3), 393–410.

Stinchcombe, A. L. (2000). Social structure and organizations, In *Economics meets sociology in strategic management*. Emerald Group Publishing Limited.

Torres, A. P., Marshall, M. I., & Sydnor, S. (2018). Does social capital pay off? the case of small business resilience after hurricane katrina. *Journal of Contingencies and Crisis Management*.

Tsai, Y.-H., Yang, C.-H. Et al. (2018). Does human capital matter to vendor profitability: Evidence from taiwan. *Journal of Economics and Management*, *14*(1), 83–106.

Tuyon, J., Mohammad, S., Junaidi, S., & Ali, R. (2011). The role of microfinance in development of micro enterprises in malaysia. *Business and Management Quarterly Review*, *2*(3), 47–57.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly*, 35–67.

Vasconcelos, F. C., & Ramirez, R. (2011). Complexity in business environments. *Journal of Business Research*, *64*(3), 236–241.

Vermeulen, B., & Pyka, A. (2017). The role of network topology and the spatial distribution and structure of knowledge in regional innovation policy: A calibrated agent-based model study. *Computational Economics*, 1–36.

Wachowski, L., & Wachowski, L. (1999). The matrix [W. B. and Roadshow E.].

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind, In *Theories of group behavior*. Springer.

Weisbuch, G. (1991). *Complex systems dynamics: An introduction to automata networks*. CRC Press.

Wolfram, S. (2018). *Cellular automata and complexity: Collected papers*. CRC Press.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The knowledge engineering review*, *10*(2), 115–152.

Xu, Y. (2016). Entrepreneurial social capital, cognitive orientation and new venture innovation. *Management Research Review*, *39*(5), 498–520.