

# Essays on Credit Access and Household Finance

**Seyma Kalay**

Doctor of Philosophy in Economics  
(Thesis Defense)

Supervisors:

Prof. Michela Cameletti  
Prof. Federica Maria Origo

**February 28, 2022**



**UNIVERSITÀ  
DEGLI STUDI  
DI BERGAMO**

**ECONOMIA**

Dipartimento di Scienze Economiche

# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Motivation

- Access to finance is defined as access to formal financial services (e.g., formal account and saving) at an affordable cost.
- It has been associated with the stability of the economy and well-being.
- Despite the advantages, it is far below the universal level particularly in developing countries.
- For a constant increase in economic growth, government's aim should be toward economically disadvantaged groups to promote them.

# Motivation

- Access to finance is defined as access to formal financial services (e.g., formal account and saving) at an affordable cost.
- It has been associated with the stability of the economy and well-being.
- Despite the advantages, it is far below the universal level particularly in developing countries.
- For a constant increase in economic growth, government's aim should be toward economically disadvantaged groups to promote them.

# Motivation

- Access to finance is defined as access to formal financial services (e.g., formal account and saving) at an affordable cost.
- It has been associated with the stability of the economy and well-being.
- Despite the advantages, it is far below the universal level particularly in developing countries.
- For a constant increase in economic growth, government's aim should be toward economically disadvantaged groups to promote them.

# Motivation

- Access to finance is defined as access to formal financial services (e.g., formal account and saving) at an affordable cost.
- It has been associated with the stability of the economy and well-being.
- Despite the advantages, it is far below the universal level particularly in developing countries.
- For a constant increase in economic growth, government's aim should be toward economically disadvantaged groups to promote them.

# Overview & Contributions

## ❶ Chapter 2

- Reviews the literature on the determinants of finance, using bibliometric techniques.

## ❷ Chapter 3

- Examines the characteristics of households to access to finance.

## ❸ Chapter 4

- Clusters the households based on their financial strength.



# Overview & Contributions

## ❶ Chapter 2

- Reviews the literature on the determinants of finance, using bibliometric techniques.

## ❷ Chapter 3

- Examines the characteristics of households to access to finance.

## ❸ Chapter 4

- Clusters the households based on their financial strength.

# Overview & Contributions

## ❶ Chapter 2

- Reviews the literature on the determinants of finance, using bibliometric techniques.

## ❷ Chapter 3

- Examines the characteristics of households to access to finance.

## ❸ Chapter 4

- Clusters the households based on their financial strength.

# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Motivation

- The literature review<sup>1</sup> plays a crucial role in gathering knowledge and guiding the future research directions (Cropanzano, 2009; Kunisch et al., 2018), regardless of discipline.
- The amount of publications is gradually increasing and it is becoming difficult to identify current trends (Aria and Cuccurullo, 2017).
- Bibliometric analysis helps to understand the scientific production, intellectual networks, trends and publication patterns between the scholars, institutions and countries (Liu, 2014; Pinto, 2014; Bourdieu, 1994; Broadus, 1987; Pritchard, 1969).
- It demonstrates the “big picture” of research (Crane, 1972) and conducts reproducible literature review concept (Broadus, 1987; Diodato and Gellatly, 2013; Pritchard et al., 1969).

---

<sup>1</sup>Depending on the purpose of the researcher (shown in Appendix 2) all types of literature reviews can be helpful (Snyder, 2019).

# Motivation

- The literature review<sup>1</sup> plays a crucial role in gathering knowledge and guiding the future research directions (Cropanzano, 2009; Kunisch et al., 2018), regardless of discipline.
- The amount of publications is gradually increasing and it is becoming difficult to identify current trends (Aria and Cuccurullo, 2017).
- Bibliometric analysis helps to understand the scientific production, intellectual networks, trends and publication patterns between the scholars, institutions and countries (Liu, 2014; Pinto, 2014; Bourdieu, 1994; Broadus, 1987; Pritchard, 1969).
- It demonstrates the “big picture” of research (Crane, 1972) and conducts reproducible literature review concept (Broadus, 1987; Diodato and Gellatly, 2013; Pritchard et al., 1969).

---

<sup>1</sup>Depending on the purpose of the researcher (shown in Appendix 2) all types of literature reviews can be helpful (Snyder, 2019).

# Motivation

- The literature review<sup>1</sup> plays a crucial role in gathering knowledge and guiding the future research directions (Cropanzano, 2009; Kunisch et al., 2018), regardless of discipline.
- The amount of publications is gradually increasing and it is becoming difficult to identify current trends (Aria and Cuccurullo, 2017).
- Bibliometric analysis helps to understand the scientific production, intellectual networks, trends and publication patterns between the scholars, institutions and countries (Liu, 2014; Pinto, 2014; Bourdieu, 1994; Broadus, 1987; Pritchard, 1969).
- It demonstrates the “big picture” of research (Crane, 1972) and conducts reproducible literature review concept (Broadus, 1987; Diodato and Gellatly, 2013; Pritchard et al., 1969).

---

<sup>1</sup>Depending on the purpose of the researcher (shown in Appendix 2) all types of literature reviews can be helpful (Snyder, 2019).

# Motivation

- The literature review<sup>1</sup> plays a crucial role in gathering knowledge and guiding the future research directions (Cropanzano, 2009; Kunisch et al., 2018), regardless of discipline.
- The amount of publications is gradually increasing and it is becoming difficult to identify current trends (Aria and Cuccurullo, 2017).
- Bibliometric analysis helps to understand the scientific production, intellectual networks, trends and publication patterns between the scholars, institutions and countries (Liu, 2014; Pinto, 2014; Bourdieu, 1994; Broadus, 1987; Pritchard, 1969).
- It demonstrates the “big picture” of research (Crane, 1972) and conducts reproducible literature review concept (Broadus, 1987; Diodato and Gellatly, 2013; Pritchard et al., 1969).

---

<sup>1</sup>Depending on the purpose of the researcher (shown in Appendix 2) all types of literature reviews can be helpful (Snyder, 2019).

# Workflow

## ❶ Step 1: Objectives of the Study

- What are the influential features under the “determinants of finance”? (such as; influential journals, articles .etc)

## ❷ Step 2: Methodology

- ❶ **Study Design:** Deciding the research question.
- ❷ **Data Collection:** Selecting a bibliometric database (e.g., WoS, Scopus, etc).
  - Selected articles based on the keyword search (e.g., determinants of finance/credit, access to credit, .etc).
  - To decide if “determinants of finance” is not marginally mentioned but a direct content in the article.
  - Finally, 210 articles were selected to conduct bibliometric analysis.
- ❸ **Data Analysis:** .txt file must be converted into a bibliographic data frame.
- ❹ **Data Visualization:** Bibliographic data frame pave the way to network matrix then, network mapping can be conducted.
- ❺ **Interpretation:** Helps researchers to make sense of bibliometric’s results.



# Workflow

## ① Step 1: Objectives of the Study

- What are the influential features under the “determinants of finance”? (such as; influential journals, articles .etc)

## ② Step 2: Methodology

- ① **Study Design:** Deciding the research question.
- ② **Data Collection:** Selecting a bibliometric database (e.g., WoS, Scopus, etc).
  - Selected articles based on the keyword search (e.g., determinants of finance/credit, access to credit, .etc).
  - To decide if “determinants of finance” is not marginally mentioned but a direct content in the article.
  - Finally, 210 articles were selected to conduct bibliometric analysis.
- ③ **Data Analysis:** .txt file must be converted into a bibliographic data frame.
- ④ **Data Visualization:** Bibliographic data frame pave the way to network matrix then, network mapping can be conducted.
- ⑤ **Interpretation:** Helps researchers to make sense of bibliometric’s results.

# Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document  $\times$  Cited reference*, *Document  $\times$  Author*, and *Document  $\times$  Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

# Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document  $\times$  Cited reference*, *Document  $\times$  Author*, and *Document  $\times$  Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

# Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document  $\times$  Cited reference*, *Document  $\times$  Author*, and *Document  $\times$  Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

## Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document  $\times$  Cited reference*, *Document  $\times$  Author*, and *Document  $\times$  Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

## Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document  $\times$  Cited reference*, *Document  $\times$  Author*, and *Document  $\times$  Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

## Bibliometric Coupling

- Documents are connected through a document's attributes (e.g., author, affiliations, cited references, keywords, etc.).
- The document-attribute matrix denoted by  $X$ , where each row is a document ( $D$ ) and each column an attribute ( $A$ ).
- The generic element of matrix  $X$  is  $x_{ij} = D_i A_j = 1$  if the  $i$ -th document has the  $j$ -th attribute, otherwise  $x_{ij} = 0$ .
- Two articles are bibliographically coupled when there is at least one commonly cited source in the reference lists of both documents, and this relationship gets stronger when the common cited sources increase (Kessler, 1963).
- The general formula for bibliometric coupling matrix can be written as

$$B_{coup} = XX^T$$

- Co-citation, Co-authorship, and Co-word can be presented as

$$B_{co} = X^T X$$

where  $X$  is a *Document*  $\times$  *Cited reference*, *Document*  $\times$  *Author*, and *Document*  $\times$  *Word* Matrix. The generic element of matrix  $B$  is  $b_{ij}$  shows the number of *co-citation*, *collaborations*, and *co-occurrence words* between documents  $i$  and  $j$ , respectively.

# Data Analysis

## ● Influential Countries

- The top three countries based on the number of published articles are USA (83), UK (26), and China (14).
- Yet, based on the total global citation per year (TGC/t) by 1 article, Netherlands (TGC/t = 31.25) and Germany (TGC/t = 19.6) are more appreciated once.

## ● Influential Affiliations

- World Bank has published many articles (14).
- Yet, Harvard University (TGC/t = 28.87) is more appreciated.

## ● Influential Journals

- World Development has published many articles (14).
- Yet, Management Science (TGC/t = 62.71), Quarterly Journal Of Economics (TGC/t = 54.12), Annual Review Of Sociology (TGC/t = 53), and Journal Of Financial Economics (TGC/t = 35.27).

## ● Influential Authors

- Beck T.(5) and Wyly Ek.(3) have the highest number of article in the data-set.
- Yet, based on the total global citation per year (TGC/t) by 1 article, Fernandes D. (TGC/t = 62.71), Khwaja Ai.(TGC/t = 54.12), Pager D.(TGC/t = 53), and Campbell Jy. (TGC/t = 51.27) are the most appreciated authors.



# Data Analysis

## ● Influential Countries

- The top three countries based on the number of published articles are USA (83), UK (26), and China (14).
- Yet, based on the total global citation per year (TGC/t) by 1 article, Netherlands (TGC/t = 31.25) and Germany (TGC/t = 19.6) are more appreciated once.

## ● Influential Affiliations

- World Bank has published many articles (14).
- Yet, Harvard University (TGC/t = 28.87) is more appreciated.

## ● Influential Journals

- World Development has published many articles (14).
- Yet, Management Science (TGC/t = 62.71), Quarterly Journal Of Economics (TGC/t = 54.12), Annual Review Of Sociology (TGC/t = 53), and Journal Of Financial Economics (TGC/t = 35.27).

## ● Influential Authors

- Beck T.(5) and Wyly Ek.(3) have the highest number of article in the data-set.
- Yet, based on the total global citation per year (TGC/t) by 1 article, Fernandes D. (TGC/t = 62.71), Khwaja Ai.(TGC/t = 54.12), Pager D.(TGC/t = 53), and Campbell Jy. (TGC/t = 51.27) are the most appreciated authors.

# Data Analysis

## ● Influential Countries

- The top three countries based on the number of published articles are USA (83), UK (26), and China (14).
- Yet, based on the total global citation per year (TGC/t) by 1 article, Netherlands (TGC/t = 31.25) and Germany (TGC/t = 19.6) are more appreciated once.

## ● Influential Affiliations

- World Bank has published many articles (14).
- Yet, Harvard University (TGC/t = 28.87) is more appreciated.

## ● Influential Journals

- World Development has published many articles (14).
- Yet, Management Science (TGC/t = 62.71), Quarterly Journal Of Economics (TGC/t = 54.12), Annual Review Of Sociology (TGC/t = 53), and Journal Of Financial Economics (TGC/t = 35.27).

## ● Influential Authors

- Beck T.(5) and Wyly Ek.(3) have the highest number of article in the data-set.
- Yet, based on the total global citation per year (TGC/t) by 1 article, Fernandes D. (TGC/t = 62.71), Khwaja Ai.(TGC/t = 54.12), Pager D.(TGC/t = 53), and Campbell Jy. (TGC/t = 51.27) are the most appreciated authors.

# Data Analysis

## ● Influential Countries

- The top three countries based on the number of published articles are USA (83), UK (26), and China (14).
- Yet, based on the total global citation per year (TGC/t) by 1 article, Netherlands (TGC/t = 31.25) and Germany (TGC/t = 19.6) are more appreciated once.

## ● Influential Affiliations

- World Bank has published many articles (14).
- Yet, Harvard University (TGC/t = 28.87) is more appreciated.

## ● Influential Journals

- World Development has published many articles (14).
- Yet, Management Science (TGC/t = 62.71), Quarterly Journal Of Economics (TGC/t = 54.12), Annual Review Of Sociology (TGC/t = 53), and Journal Of Financial Economics (TGC/t = 35.27).

## ● Influential Authors

- Beck T.(5) and Wyly Ek.(3) have the highest number of article in the data-set.
- Yet, based on the total global citation per year (TGC/t) by 1 article, Fernandes D. (TGC/t = 62.71), Khwaja Ai.(TGC/t = 54.12), Pager D.(TGC/t = 53), and Campbell Jy. (TGC/t = 51.27) are the most appreciated authors.

# Data Visualization

## • Co-Citation

- The literature is divided into two main streams: (i) Lending to small and (ii) Lending to big borrowers.

## • Co-Word

- Co-words network prints out the keywords for each streams.
- (i): Financial inclusion, financial literacy, financial development, financial institutions, household finance, race, credit, micro finance, India, gender, discrimination, entrepreneurship, and financial constraints. (ii): SMEs, credit constraints, banking, formal credit, political connections, China, and social capital.

## • Co-Authorship

- Co-authorship network shows that Bect T. and Cull R. have a larger collaboration network between the authors.
- Demirguc-Kunt A. has a direct relationship with Beck T., Klapper L., and Allen F..
- Cull R. and Bect T. seem to have a higher quantity of articles comparing the other authors in the network cluster.

# Data Visualization

## ● Co-Citation

- The literature is divided into two main streams: (i) Lending to small and (ii) Lending to big borrowers.

## ● Co-Word

- Co-words network prints out the keywords for each streams.
- (i): Financial inclusion, financial literacy, financial development, financial institutions, household finance, race, credit, micro finance, India, gender, discrimination, entrepreneurship, and financial constraints. (ii): SMEs, credit constraints, banking, formal credit, political connections, China, and social capital.

## ● Co-Authorship

- Co-authorship network shows that Bect T. and Cull R. have a larger collaboration network between the authors.
- Demirguc-Kunt A. has a direct relationship with Beck T., Klapper L., and Allen F..
- Cull R. and Bect T. seem to have a higher quantity of articles comparing the other authors in the network cluster.

# Data Visualization

## • Co-Citation

- The literature is divided into two main streams: (i) Lending to small and (ii) Lending to big borrowers.

## • Co-Word

- Co-words network prints out the keywords for each streams.
- (i): Financial inclusion, financial literacy, financial development, financial institutions, household finance, race, credit, micro finance, India, gender, discrimination, entrepreneurship, and financial constraints. (ii): SMEs, credit constraints, banking, formal credit, political connections, China, and social capital.

## • Co-Authorship

- Co-authorship network shows that Bect T. and Cull R. have a larger collaboration network between the authors.
- Demirguc-Kunt A. has a direct relationship with Beck T., Klapper L., and Allen F..
- Cull R. and Bect T. seem to have a higher quantity of articles comparing the other authors in the network cluster.

# Data Visualization

## ● Co-Citation

- The literature is divided into two main streams: (i) Lending to small and (ii) Lending to big borrowers.

## ● Co-Word









- Co-words network prints out the keywords for each streams.
- (i): Financial inclusion, financial literacy, financial development, financial institutions, household finance, race, credit, micro finance, India, gender, discrimination, entrepreneurship, and financial constraints. (ii): SMEs, credit constraints, banking, formal credit, political connections, China, and social capital.

## ● Co-Authorship

- Co-authorship network shows that Bect T. and Cull R. have a larger collaboration network between the authors.
- Demirguc-Kunt A. has a direct relationship with Beck T., Klapper I., and Allen F..
- Cull R. and Bect T. seem to have a higher quantity of articles comparing the other authors in the network cluster.

# Biblio App

Biblio is a [Shinyapp](#)<sup>2</sup> which provides an interface for bibliometric analysis, where the results of this study can be reproducible.

- ❶ **Study Design:** Deciding a research question.
- ❷ **Data Collection:** Collecting the data-set<sup>3</sup>. Users should wait until "Upload Complete" then click "Start Conversion".
- ❸ **Data Analysis:** Descriptive statistics can be seen under the  Data,  Authors, and  Citations.
- ❹ **Data Visualization:** Network analysis can be seen under the  Tree,  Map,  Words,  Thematic Map, and  Network tabs.

---

<sup>2</sup>See at: <https://seymakalay87.shinyapps.io/biblio/>

<sup>3</sup>The app default data-set is a .txt file which can be retrieved from WoS database. For Scopus database users select "Load bibliometrix file(s)" and upload the .bib file.



# Conclusion

- ❶ Defining the influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
- ❷ Identifying the main research streams through co-citation and co-word analysis.
- ❸ Network analysis between Co-citation, Co-author, and Co-word.
- ❹ Identifying 13 research questions, after combining quantitative (bibliometric) and qualitative (content) analysis.
- ❺ A Shinyapp for the reproducible future studies see at:  
<https://seymakalay87.shinyapps.io/biblio/>.

# Conclusion

- 1 Defining the influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
- 2 Identifying the main research streams through co-citation and co-word analysis.
- 3 Network analysis between Co-citation, Co-author, and Co-word.
- 4 Identifying 13 research questions, after combining quantitative (bibliometric) and qualitative (content) analysis.
- 5 A Shinyapp for the reproducible future studies see at:  
<https://seymakalay87.shinyapps.io/biblio/>.

# Conclusion

- ❶ Defining the influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
- ❷ Identifying the main research streams through co-citation and co-word analysis.
- ❸ Network analysis between Co-citation, Co-author, and Co-word.
- ❹ Identifying 13 research questions, after combining quantitative (bibliometric) and qualitative (content) analysis.
- ❺ A Shinyapp for the reproducible future studies see at:  
<https://seymakalay87.shinyapps.io/biblio/>.

# Conclusion

- ❶ Defining the influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
- ❷ Identifying the main research streams through co-citation and co-word analysis.
- ❸ Network analysis between Co-citation, Co-author, and Co-word.
- ❹ Identifying 13 research questions, after combining quantitative (bibliometric) and qualitative (content) analysis.
- ❺ A Shinyapp for the reproducible future studies see at:  
<https://seymakalay87.shinyapps.io/biblio/>.

# Conclusion

- ❶ Defining the influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
- ❷ Identifying the main research streams through co-citation and co-word analysis.
- ❸ Network analysis between Co-citation, Co-author, and Co-word.
- ❹ Identifying 13 research questions, after combining quantitative (bibliometric) and qualitative (content) analysis.
- ❺ A Shinyapp for the reproducible future studies see at:  
<https://seymakalay87.shinyapps.io/biblio/>.

# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Motivation

- China has the world's largest economy and has experienced a rapid economic growth over the past few decades.
- Yet, China is still considered a developing country, and millions are below the international poverty standards.
- The use of formal financial services is far lower than other emerging (Fungacova et al., 2014) and high-income economies (Demirguc-Kunt and Klapper, 2012).

# Motivation

- China has the world's largest economy and has experienced a rapid economic growth over the past few decades.
- Yet, China is still considered a developing country, and millions are below the international poverty standards.
- The use of formal financial services is far lower than other emerging (Fungacova et al., 2014) and high-income economies (Demirguc-Kunt and Klapper, 2012).



# Motivation

- China has the world's largest economy and has experienced a rapid economic growth over the past few decades.
- Yet, China is still considered a developing country, and millions are below the international poverty standards.
- The use of formal financial services is far lower than other emerging (Fungacova et al., 2014) and high-income economies (Demirguc-Kunt and Klapper, 2012).

# Workflow

## ① Step 1: Objectives of the Study

- To understand the characteristics of Chinese households to access to credit and its type.

## ② Step 2: Methodology

- I split the CHFS data-set<sup>4</sup> into 4 different data-splits<sup>5</sup> (Urb & Rrl, Educ.0 & Educ.1, CCP.0 & CCP.1, and Sex.0 & Sex.1).
- Using each time one of the asset owning variables (net-worth, NW-HE, and liquid assets) interchangeable on the each data set.
- In total, I built, 120 models, 30 linear and 90 ML models, to explain the characteristics of Chinese households for both access to loan and its type.
- Each data-set was split into 80:20 train:test groups and 10-CV were applied.
- **Binary Regression:** Access to loan
  - Response variable: "Access to Loan" = 1, and "Otherwise" = 0
- **Multi Regression:** Access to loan type
  - Response variable: "Formal Loan" = 1, "Informal Loan" = 2, "Both Loans" = 3, and "Otherwise" = 0

<sup>4</sup>In this study, "CHFS", "Benchmark", and "BchMk" are used interchangeably.

<sup>5</sup>See the abbreviations in Appendix 3.

# Workflow

## ① Step 1: Objectives of the Study

- To understand the characteristics of Chinese households to access to credit and its type.

## ② Step 2: Methodology

- I split the CHFS data-set<sup>4</sup> into 4 different data-splits<sup>5</sup> (Urb & Rrl, Educ.0 & Educ.1, CCP.0 & CCP.1, and Sex.0 & Sex.1).
- Using each time one of the asset owning variables (net-worth, NW-HE, and liquid assets) interchangeable on the each data set.
- In total, I built, 120 models, 30 linear and 90 ML models, to explain the characteristics of Chinese households for both access to loan and its type.
- Each data-set was split into 80:20 train:test groups and 10-CV were applied.
- **Binary Regression:** Access to loan
  - Response variable: "Access to Loan" = 1, and "Otherwise" = 0
- **Multi Regression:** Access to loan type
  - Response variable: "Formal Loan" = 1, "Informal Loan" = 2, "Both Loans" = 3, and "Otherwise" = 0

---

<sup>4</sup>In this study, "CHFS", "Benchmark", and "BchMk" are used interchangeably.

<sup>5</sup>See the abbreviations in Appendix 3.

# Workflow

## ① Step 1: Objectives of the Study

- To understand the characteristics of Chinese households to access to credit and its type.

## ② Step 2: Methodology

- I split the CHFS data-set<sup>4</sup> into 4 different data-splits<sup>5</sup> (Urb & Rrl, Educ.0 & Educ.1, CCP.0 & CCP.1, and Sex.0 & Sex.1).
- Using each time one of the asset owning variables (net-worth, NW-HE, and liquid assets) interchangeable on the each data set.
- In total, I built, 120 models, 30 linear and 90 ML models, to explain the characteristics of Chinese households for both access to loan and its type.
- Each data-set was split into 80:20 train:test groups and 10-CV were applied.
- **Binary Regression:** Access to loan
  - Response variable: "Access to Loan" = 1, and "Otherwise" = 0
- **Multi Regression:** Access to loan type
  - Response variable: "Formal Loan" = 1, "Informal Loan" = 2, "Both Loans" = 3, and "Otherwise" = 0

---

<sup>4</sup>In this study, "CHFS", "Benchmark", and "BchMk" are used interchangeably.

<sup>5</sup>See the abbreviations in Appendix 3.

# Classification Models

## • Generalized Logistic Regression (GLM)

- Let  $\mathbf{y}$  be a vector values for the response variable of accessing credit for each applicant  $n$ , such that  $y_i = 1$  if the applicant- $i$  has access to credit, and zero otherwise. Furthermore, let  $\mathbf{x} = \{x_{i,j}\}$ , where  $i = 1, \dots, n$  and  $j = 1, \dots, p$  be a matrix of  $n$  observations with  $p$  characteristics of the applicants. The log-odds ratio can be defined as

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \mathbf{x}_i \beta = \beta_0 + \sum_{j=1}^p \beta_j x_i \quad (1)$$

where  $\pi_i = P(y_i = 1 | \mathbf{x}_i)$ ,  $\beta_0$  is the intercept,  $\beta = (\beta_1, \dots, \beta_p)'$  is a  $p \times 1$  vector of coefficients and  $\mathbf{x}_i$  is the  $i$ -th row of  $\mathbf{x}$ .

## • Multinomial Logistic Model (MLM)

- Multi-nominal model is the generalized form of binary logistic model (1) and can be defined as

$$\pi_i^h = P(y_i^h = 1 | \mathbf{x}_i^h) \quad (2)$$

where  $h$  presents the class labels ("1-of- $h$ ") on the basis of an input vector  $\mathbf{x}_i$ .

- Furthermore,  $y_i^h = 1$  if the weight  $\mathbf{w}$  of  $\mathbf{x}_i$  corresponds to belong a class and  $y_i^h = 0$  otherwise.

The weight vectors  $\mathbf{w}^j$  for  $i \in \{1, \dots, h-1\}$ , and the class probabilities must satisfy

$$\sum_{i=1}^h P(y_i^h = 1 | \mathbf{x}_i^h, \mathbf{w}) = 1$$

# Classification Models

## • Generalized Logistic Regression (GLM)

- Let  $\mathbf{y}$  be a vector values for the response variable of accessing credit for each applicant  $n$ , such that  $y_i = 1$  if the applicant- $i$  has access to credit, and zero otherwise. Furthermore, let  $\mathbf{x} = \{x_{i,j}\}$ , where  $i = 1, \dots, n$  and  $j = 1, \dots, p$  be a matrix of  $n$  observations with  $p$  characteristics of the applicants. The log-odds ratio can be defined as

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \mathbf{x}_i \beta = \beta_0 + \sum_{j=1}^p \beta_j x_i \quad (1)$$

where  $\pi_i = P(y_i = 1 | \mathbf{x}_i)$ ,  $\beta_0$  is the intercept,  $\beta = (\beta_1, \dots, \beta_p)'$  is a  $p \times 1$  vector of coefficients and  $\mathbf{x}_i$  is the  $i$ -th row of  $\mathbf{x}$ .

## • Multinomial Logistic Model (MLM)

- Multi-nominal model is the generalized form of binary logistic model (1) and can be defined as

$$\pi_i^h = P(y_i^h = 1 | \mathbf{x}_i^h) \quad (2)$$

where  $h$  presents the class labels ("1-of- $h$ ") on the basis of an input vector  $\mathbf{x}_i$ .

- Furthermore,  $y_i^h = 1$  if the weight  $\mathbf{w}$  of  $\mathbf{x}_i$  corresponds to belong a class and  $y_i^h = 0$  otherwise.

The weight vectors  $\mathbf{w}^i$  for  $i \in \{1, \dots, h-1\}$ , and the class probabilities must satisfy

$$\sum_{i=1}^h P(y_i^h = 1 | \mathbf{x}_i^h, \mathbf{w}) = 1$$

# Tree Models for Robustness Check

## • Bagging (BAG)

- Generating  $B$  different bootstrapped training data-sets  $(\hat{f}^{*1}(x), \hat{f}^{*2}(x), \dots, \hat{f}^{*B}(x))$  and then average the resulting predictions

$$\hat{f}_{avg}(x) = \frac{1}{B} \sum_{i=1}^B \hat{f}^{*b}(x) \quad (3)$$

## • Random Forest (RF)

- Random forests<sup>6</sup> which de-correlate the trees by considering  $m_{try} \approx \sqrt{p}$  show an improvement over bagged trees where  $m = p$  (Breiman, 1984).

## • Gradient Boosting (BOOST)

- Each tree is fit using information from previous trees.

$$\hat{\pi}_i = \frac{1}{1 + \exp[-f(x)]} \quad (4)$$

where  $f(x)$  is a model prediction in the range of  $[-\infty, \infty]$  and its initial estimate of the model is  $\hat{f}_i^{(0)} = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ , where  $\hat{\pi}$  is the estimated sample proportion of a single class from the training set.

---

<sup>6</sup> Random forests' tuning parameter is the number of randomly selected predictors,  $p$ , to choose from at each split, and is commonly referred to as  $m_{try}$ .

# Tree Models for Robustness Check

## • Bagging (BAG)

- Generating  $B$  different bootstrapped training data-sets  $(\hat{f}^{*1}(x), \hat{f}^{*2}(x), \dots, \hat{f}^{*B}(x))$  and then average the resulting predictions

$$\hat{f}_{avg}(x) = \frac{1}{B} \sum_{i=1}^B \hat{f}^{*b}(x) \quad (3)$$

## • Random Forest (RF)

- Random forests<sup>6</sup> which de-correlate the trees by considering  $m_{try} \approx \sqrt{p}$  show an improvement over bagged trees where  $m = p$  (Breiman, 1984).

## • Gradient Boosting (BOOST)

- Each tree is fit using information from previous trees.

$$\hat{\pi}_i = \frac{1}{1 + \exp[-f(x)]} \quad (4)$$

where  $f(x)$  is a model prediction in the range of  $[-\infty, \infty]$  and its initial estimate of the model is  $\hat{f}_i^{(0)} = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ , where  $\hat{\pi}$  is the estimated sample proportion of a single class from the training set.

---

<sup>6</sup>Random forests' tuning parameter is the number of randomly selected predictors,  $p$ , to choose from at each split, and is commonly referred to as  $m_{try}$ .



# Tree Models for Robustness Check

## • Bagging (BAG)

- Generating  $B$  different bootstrapped training data-sets  $(\hat{f}^{*1}(x), \hat{f}^{*2}(x), \dots, \hat{f}^{*B}(x))$  and then average the resulting predictions

$$\hat{f}_{avg}(x) = \frac{1}{B} \sum_{i=1}^B \hat{f}^{*i}(x) \quad (3)$$

## • Random Forest (RF)

- Random forests<sup>6</sup> which de-correlate the trees by considering  $m_{try} \approx \sqrt{p}$  show an improvement over bagged trees where  $m = p$  (Breiman, 1984).

## • Gradient Boosting (BOOST)

- Each tree is fit using information from previous trees.

$$\hat{\pi}_i = \frac{1}{1 + \exp[-f(x)]} \quad (4)$$

where  $f(x)$  is a model prediction in the range of  $[-\infty, \infty]$  and its initial estimate of the model is  $f_i^{(0)} = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ , where  $\hat{\pi}$  is the estimated sample proportion of a single class from the training set.

---

<sup>6</sup>Random forests' tuning parameter is the number of randomly selected predictors,  $p$ , to choose from at each split, and is commonly referred to as  $m_{try}$ .

# Results of Classification Models: AUC

| Data splits     | <i>y = Access Loan</i> |       |       | <i>y = Loan Type</i> |       |       |
|-----------------|------------------------|-------|-------|----------------------|-------|-------|
|                 | 1                      | 2     | 3     | 1                    | 2     | 3     |
| GLM             |                        |       | MLM   |                      |       |       |
| Urb & Rrl       | 0.699                  | 0.696 | 0.699 | 0.710                | 0.707 | 0.710 |
| Educ.0 & Educ.1 | 0.685                  | 0.683 | 0.686 | 0.712                | 0.708 | 0.712 |
| CCP.0 & CCP.1   | 0.708                  | 0.706 | 0.708 | 0.720                | 0.717 | 0.720 |
| SEX.0 & SEX.1   | 0.680                  | 0.677 | 0.681 | 0.705                | 0.702 | 0.706 |
| BchMk           | 0.698                  | 0.695 | 0.699 | 0.712                | 0.709 | 0.709 |
| BAG             |                        |       | BAG   |                      |       |       |
| Urb & Rrl       | 0.668                  | 0.661 | 0.663 | 0.664                | 0.658 | 0.659 |
| Educ.0 & Educ.1 | 0.662                  | 0.655 | 0.644 | 0.671                | 0.676 | 0.668 |
| CCP.0 & CCP.1   | 0.664                  | 0.667 | 0.662 | 0.676                | 0.676 | 0.676 |
| SEX.0 & SEX.1   | 0.659                  | 0.662 | 0.653 | 0.671                | 0.666 | 0.676 |
| BchMk           | 0.667                  | 0.664 | 0.660 | 0.669                | 0.677 | 0.666 |
| RF              |                        |       | RF    |                      |       |       |
| Urb & Rrl       | 0.688                  | 0.687 | 0.685 | 0.677                | 0.674 | 0.680 |
| Educ.0 & Educ.1 | 0.672                  | 0.669 | 0.668 | 0.679                | 0.678 | 0.680 |
| CCP.0 & CCP.1   | 0.691                  | 0.690 | 0.690 | 0.686                | 0.685 | 0.690 |
| SEX.0 & SEX.1   | 0.670                  | 0.672 | 0.664 | 0.683                | 0.675 | 0.679 |
| BchMk           | 0.687                  | 0.683 | 0.682 | 0.689                | 0.682 | 0.687 |
| GBM             |                        |       | GBM   |                      |       |       |
| Urb & Rrl       | 0.718                  | 0.722 | 0.716 | 0.721                | 0.719 | 0.722 |
| Educ.0 & Educ.1 | 0.700                  | 0.701 | 0.700 | 0.726                | 0.722 | 0.726 |
| CCP.0 & CCP.1   | 0.721                  | 0.725 | 0.718 | 0.732                | 0.733 | 0.733 |
| SEX.0 & SEX.1   | 0.699                  | 0.700 | 0.694 | 0.722                | 0.720 | 0.724 |
| BchMk           | 0.717                  | 0.718 | 0.722 | 0.725                | 0.725 | 0.729 |

Note: AUC of the Model (1:3) presents Network, NW-HE, and Liquid Assets, as predictor, respectively. The definition of the variables can be found in Appendix 3.

Generalized Logistic Models ( $y = \text{Access Loan}$ )

| Variables      | GLM.1   |         |         | GLM.2   |         |         | GLM.3   |         |         |
|----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|                | BchMk   | CCP.0   | CCP.1   | BchMk   | CCP.0   | CCP.1   | BchMk   | CCP.0   | CCP.1   |
| Gender         | 0.95    | 0.97    | 0.97    | 0.94*   | 0.97    | 0.96    | 0.95    | 0.97    | 0.97    |
| Marital Status | 1.25*** | 1.32*** | 1.36**  | 1.27*** | 1.33*** | 1.39**  | 1.25*** | 1.32*** | 1.36**  |
| Age            | 0.58*** | 0.62*** | 0.43*** | 0.58*** | 0.62*** | 0.43*** | 0.58*** | 0.62*** | 0.43*** |
| Employed       | 1.27*** | 1.17*** | 1.40*** | 1.26*** | 1.17*** | 1.39*** | 1.27*** | 1.17*** | 1.40*** |
| Education      | 1.14*** | 1.10**  | 1.22**  | 1.16*** | 1.11*** | 1.25*** | 1.14*** | 1.09**  | 1.21**  |
| Party          | 1.05    | —       | —       | 1.06    | —       | —       | 1.05    | —       | —       |
| HR             | 0.75*** | 0.75*** | 0.78*** | 0.76*** | 0.76*** | 0.80**  | 0.74*** | 0.75*** | 0.78*** |
| Region-East    | 0.64*** | 0.65*** | 0.72*** | 0.66*** | 0.66*** | 0.75*** | 0.64*** | 0.65*** | 0.72*** |
| Region-Center  | 0.84*** | 0.86*** | 0.92    | 0.84*** | 0.86*** | 0.91    | 0.84*** | 0.86*** | 0.92    |
| Fin.Inter      | 1.02    | 1.04    | 1.01    | 1.03    | 1.05    | 1.02    | 1.02    | 1.04    | 1.01    |
| Fin.Knowledge  | 1.70*** | 1.77*** | 1.39*** | 1.74*** | 1.81*** | 1.42*** | 1.69*** | 1.76*** | 1.39*** |
| Income         | 1.12*** | 1.10*** | 1.16**  | 1.16*** | 1.14*** | 1.24*** | 1.11*** | 1.10*** | 1.15**  |
| Network        | 1.20*** | 1.15*** | 1.44*** | —       | —       | —       | —       | —       | —       |
| NW-HE          | —       | —       | —       | 1.10*** | 1.07*** | 1.25*** | —       | —       | —       |
| Liquid Assets  | —       | —       | —       | —       | —       | —       | 1.21*** | 1.17*** | 1.47*** |
| Observations   | 26,212  | 21,184  | 5,029   | 26,212  | 21,184  | 5,029   | 26,212  | 21,184  | 5,029   |
| Log Likelihood | -14,794 | -12,158 | -2,600  | -14,834 | -12,183 | -2,619  | -14,785 | -12,152 | -2,597  |

Note: Asset holding variables Network, NW-HE (net-worth minus home equity), and liquid assets. Odd ratios (OR) are reported with significance: \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ . The number of the observation is approximately 80% of the total observation and the total number of observation may be vary based on the data-split group. The variables and abbreviations can be found in Appendix 3.

# Multinomial Logistic Models ( $y = \text{Loan Type}$ )

| Variables      | Fml       | Infm    | Both    | Fml      | Infm    | Both    | Fml       | Infm    | Both    |
|----------------|-----------|---------|---------|----------|---------|---------|-----------|---------|---------|
| MLM.3          | BchMk     |         |         | CCP.0    |         |         | CCP.1     |         |         |
| Gender         | 0.84***   | 1.11*** | 0.97*** | 0.87***  | 1.17*** | 0.98*** | 0.85***   | 1.15*** | 0.98*** |
| Marital Status | 1.93***   | 1.15*** | 1.92*** | 1.44***  | 1.15*** | 2.50*** | 1.93***   | 1.13*** | 1.93*** |
| Age            | 0.95***   | 0.98*** | 0.95*** | 0.94***  | 0.97*** | 0.94*** | 0.95***   | 0.98*** | 0.95*** |
| Employed       | 1.61***   | 1.07*** | 1.57*** | 1.72***  | 1.19*** | 2.31*** | 1.58***   | 1.00*** | 1.37*** |
| Education      | 2.07***   | 0.62*** | 1.16*** | 1.96***  | 0.73*** | 1.28*** | 2.07***   | 0.65*** | 1.07*** |
| Party          | 1.43***   | 0.78*** | 1.13*** | —        | —       | —       | —         | —       | —       |
| HR             | 1.51***   | 0.64*** | 0.72*** | 1.76***  | 0.51*** | 0.58*** | 1.49***   | 0.66*** | 0.78*** |
| Region-East    | 0.77***   | 0.66*** | 0.56*** | 0.94***  | 0.54*** | 0.57*** | 0.77***   | 0.65*** | 0.53*** |
| Region-Center  | 0.64***   | 1.03*** | 0.76*** | 0.90***  | 0.97*** | 0.84*** | 0.60***   | 1.01*** | 0.75*** |
| Fin.Inter      | 1.22***   | 0.98*** | 1.10*** | 1.10***  | 0.77*** | 1.17*** | 1.23***   | 1.02*** | 1.08*** |
| Fin.Knowledge  | 2.07***   | 0.79*** | 2.10*** | 1.57***  | 0.57*** | 1.99*** | 2.14***   | 0.99*** | 2.30*** |
| Income         | 1.00***   | 1.00*** | 1.00*** | 1.00***  | 1.00*   | 1.00*** | 1.00***   | 1.00*** | 1.00*** |
| Liquid Assets  | 1.00***   | 1.00*** | 1.00*** | 1.00***  | 1.00**  | 1.00*** | 1.00***   | 1.00*** | 1.00*** |
| Observations   | 26212     |         |         | 21183    |         |         | 5028      |         |         |
| AIC            | 42,258.51 |         |         | 7,681.80 |         |         | 34,677.61 |         |         |

Note: Multinomial Logistic Model with Liquid Assets as Predictor. Relative Risk Ratios (RRR) are reported with significance: \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ . The number of the observation is approximately 80% of the total observation and the total number of observation may be vary based on the data-split group. The variables and abbreviations can be found in Appendix 3.

# CRAN - Package pomodoro

## 1 Step 1: Install and Call the Library

```
1 install("pomodoro")
2 library(pomodoro)
```

## 2 Step 2: Built a Selected Model:

- 3 Automatically conducts stratified random sampling on 80/20 train/test set with 10 CV and gives *finModel* results.

```
1 yvar <- c("Loan.Type") #or yvar <- c("multi.level")
2 xvar <- c("sex", "married", "age", "havejob", "educ", "political.afl",
3         "rural", "region", "fin.intermdiarie", "fin.knowldge", "income")
4 set.seed(123)
5 BchMk.MLR.1 <- RF_Model(sample_data, c(xvar, "networth"), yvar)
```

## 4 Step 3: Estimated Models:

- 6 To model all the data set and its splits, interchangeably with 3 assets owning variable.

```
1 CCP.RF <- Estimate_Models(sample_data, yvar, xvec = xvar, exog = "political.afl",
2                           xadd = c("networth", "networth_homeequity", "liquid.assets"),
3                           type = "RF", dnames = c("0", "1"))
```

## 6 Step 4: Combined Performance:

- 7 To find combined model performance.

```
1 Sub.CCP.RF <- list(Mdl.1 = CCP.RF$EstMdl$'D.1+networth',
2                   Mdl.0 = CCP.RF$EstMdl$'D.0+networth')
3 CCP.NoCCP.RF <- Combined_Performance(Sub.CCP.RF)
```

# CRAN - Package pomodoro

## ④ Step 1: Install and Call the Library

```
1 install("pomodoro")
2 library(pomodoro)
```

## ② Step 2: Built a Selected Model:

- ⑧ Automatically conducts stratified random sampling on 80/20 train/test set with 10 CV and gives *finamodel* results.

```
1 yvar <- c("Loan.Type") #or yvar <- c("multi.level")
2 xvar <- c("sex", "married", "age", "havejob", "educ", "political.afl",
3         "rural", "region", "fin.intermdiarries", "fin.knowldge", "income")
4 set.seed(123)
5 BchMk.MLR.1 <- RF_Model(sample_data, c(xvar, "networth"), yvar)
```

## ④ Step 3: Estimated Models:

- ⑥ To model all the data set and its splits, interchangeably with 3 assets owning variable.

```
1 CCP.RF <- Estimate_Models(sample_data, yvar, xvec = xvar, exog = "political.afl",
2                           xadd = c("networth", "networth_homeequity", "liquid.assets"),
3                           type = "RF", dnames = c("0", "1"))
```

## ⑥ Step 4: Combined Performance:

- ⑦ To find combined model performance.

```
1 Sub.CCP.RF <- list(Mdl.1 = CCP.RF$EstMdl$'D.1+networth',
2                   Mdl.0 = CCP.RF$EstMdl$'D.0+networth')
3 CCP.NoCCP.RF <- Combined_Performance(Sub.CCP.RF)
```

## CRAN - Package pomodoro

### ④ Step 1: Install and Call the Library

```
1 install("pomodoro")
2 library(pomodoro)
```

### ② Step 2: Built a Selected Model:

- ⑤ Automatically conducts stratified random sampling on 80/20 train/test set with 10 CV and gives *finamodel* results.

```
1 yvar <- c("Loan.Type") #or yvar <- c("multi.level")
2 xvar <- c("sex", "married", "age", "havejob", "educ", "political.afl",
3         "rural", "region", "fin.intermdiarries", "fin.knowldge", "income")
4 set.seed(123)
5 BchMk.MLR.1 <- RF_Model(sample_data, c(xvar, "networth"), yvar)
```

### ④ Step 3: Estimated Models:

- ⑤ To model all the data set and its splits, interchangeably with 3 assets owning variable.

```
1 CCP.RF <- Estimate_Models(sample_data, yvar, xvec = xvar, exog = "political.afl",
2                           xadd = c("networth", "networth_homequity", "liquid.assets"),
3                           type = "RF", dnames = c("0", "1"))
```

### ⑥ Step 4: Combined Performance:

- ⑦ To find combined model performance.

```
1 Sub.CCP.RF <- list(Mdl.1 = CCP.RF$EstMdl$'D.1+networth',
2                   Mdl.0 = CCP.RF$EstMdl$'D.0+networth')
3 CCP.NoCCP.RF <- Combined_Performance(Sub.CCP.RF)
```

# CRAN - Package pomodoro

## ④ Step 1: Install and Call the Library

```
1 install("pomodoro")
2 library(pomodoro)
```

## ② Step 2: Built a Selected Model:

- ⑤ Automatically conducts stratified random sampling on 80/20 train/test set with 10 CV and gives *finModel* results.

```
1 yvar <- c("Loan.Type") #or yvar <- c("multi.level")
2 xvar <- c("sex", "married", "age", "havejob", "educ", "political.afl",
3         "rural", "region", "fin.intermdiarries", "fin.knowldge", "income")
4 set.seed(123)
5 BchMk.MLR.1 <- RF_Model(sample_data, c(xvar, "networth"), yvar)
```

## ④ Step 3: Estimated Models:

- ⑤ To model all the data set and its splits, interchangeably with 3 assets owning variable.

```
1 CCP.RF <- Estimate_Models(sample_data, yvar, xvec = xvar, exog = "political.afl",
2                           xadd = c("networth", "networth_homequity", "liquid.assets"),
3                           type = "RF", dnames = c("0", "1"))
```

## ⑥ Step 4: Combined Performance:

- ⑦ To find combined model performance.

```
1 Sub.CCP.RF <- list(Mdl.1 = CCP.RF$EstMdl$`D.1+networth`,
2                   Mdl.0 = CCP.RF$EstMdl$`D.0+networth`)
3 CCP.NoCCP.RF <- Combined_Performance(Sub.CCP.RF)
```



# Conclusion

- 1 CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- 2 The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- 3 Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- 4 Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- 5 Additionally, increasing the number of financial intermediaries and the household financial knowledge can help to access formal loans.
- 6 A CRAN - Package pomodoro for reproducibility see at: <https://seymakalay.github.io/pomodoro/index.html>.

# Conclusion

- 1 CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- 2 The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- 3 Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- 4 Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- 5 Additionally, increasing the number of financial intermediaries and the household financial knowledge can help to access formal loans.
- 6 A CRAN - Package pomodoro for reproducibility see at: <https://seymakalay.github.io/pomodoro/index.html>.

# Conclusion

- ① CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- ② The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- ③ Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- ④ Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- ⑤ Additionally, increasing the number of financial intermediaries and the household financial knowledge can help to access formal loans.
- ⑥ A CRAN - Package pomodoro for reproducibility see at: <https://seymakalay.github.io/pomodoro/index.html>.

# Conclusion

- ④ CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- ② The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- ③ Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- ④ Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- ⑤ Additionally, increasing the number financial intermediaries and the household financial knowledge can help to access formal loans.
- ⑥ A CRAN - Package pomodoro for reproducibility see at:  
<https://seymakalay.github.io/pomodoro/index.html>.

# Conclusion

- ④ CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- ② The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- ③ Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- ④ Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- ⑤ Additionally, increasing the number of financial intermediaries and the household financial knowledge can help to access formal loans.
- ⑥ A CRAN - Package pomodoro for reproducibility see at:  
<https://seymakalay.github.io/pomodoro/index.html>.

# Conclusion

- ④ CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
- ② The findings of this study are in line with (Fungacova and Weill, 2015; Chen and Jin, 2017) and indicate credit ownership is low.
- ③ Formal financial inclusion is particularly constrained which is distributed economically advantaged groups and political affiliation can help to access loan (Faccio, 2006; Khwaja and Mian, 2005).
- ④ Policies must target economically disadvantaged households to improve the financial inclusion (Gan et al., 2014).
- ⑤ Additionally, increasing the number financial intermediates and the household financial knowledge can help to access formal loans.
- ⑥ A CRAN - Package pomodoro for reproducibility see at:  
<https://seymakalay.github.io/pomodoro/index.html>.

# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Motivation

- After the Communist Party won the civil war which lasts more than 20 years, private enterprises were fully banned in China between 1952 and 1977.
- The new economic reforms legislated in 1978, policies target the urban sector (Wan, 2008). As a result most of the formal financial providers were closed in rural/poor areas (Hannig and Jansen, 2010; Sparreboom and Duflos, 2012).
- One of the major reasons for poor remaining poor is linked to lack of access to formal credit (Collins et al., 2009) and the existence of a large gap between demand and supply (Sparreboom and Duflos, 2012).
- Understanding the barriers for accessing finance is crucial; they can be either geographic (e.g., absence of nearby bank branches) or socioeconomic (e.g., minimum income & high collateral requirements, social, or ethnic groups) (Hannig and Jansen, 2010).



# Motivation

- After the Communist Party won the civil war which lasts more than 20 years, private enterprises were fully banned in China between 1952 and 1977.
- The new economic reforms legislated in 1978, policies target the urban sector (Wan, 2008). As a result most of the formal financial providers were closed in rural/poor areas (Hannig and Jansen, 2010; Sparreboom and Duflos, 2012).
- One of the major reasons for poor remaining poor is linked to lack of access to formal credit (Collins et al., 2009) and the existence of a large gap between demand and supply (Sparreboom and Duflos, 2012).
- Understanding the barriers for accessing finance is crucial; they can be either geographic (e.g., absence of nearby bank branches) or socioeconomic (e.g., minimum income & high collateral requirements, social, or ethnic groups) (Hannig and Jansen, 2010).

# Motivation

- After the Communist Party won the civil war which lasts more than 20 years, private enterprises were fully banned in China between 1952 and 1977.
- The new economic reforms legislated in 1978, policies target the urban sector (Wan, 2008). As a result most of the formal financial providers were closed in rural/poor areas (Hannig and Jansen, 2010; Sparreboom and Duflos, 2012).
- One of the major reasons for poor remaining poor is linked to lack of access to formal credit (Collins et al., 2009) and the existence of a large gap between demand and supply (Sparreboom and Duflos, 2012).
- Understanding the barriers for accessing finance is crucial; they can be either geographic (e.g., absence of nearby bank branches) or socioeconomic (e.g., minimum income & high collateral requirements, social, or ethnic groups) (Hannig and Jansen, 2010).

# Motivation

- After the Communist Party won the civil war which lasts more than 20 years, private enterprises were fully banned in China between 1952 and 1977.
- The new economic reforms legislated in 1978, policies target the urban sector (Wan, 2008). As a result most of the formal financial providers were closed in rural/poor areas (Hannig and Jansen, 2010; Sparreboom and Duflos, 2012).
- One of the major reasons for poor remaining poor is linked to lack of access to formal credit (Collins et al., 2009) and the existence of a large gap between demand and supply (Sparreboom and Duflos, 2012).
- Understanding the barriers for accessing finance is crucial; they can be either geographic (e.g., absence of nearby bank branches) or socioeconomic (e.g., minimum income & high collateral requirements, social, or ethnic groups) (Hannig and Jansen, 2010).

# Workflow

## ① Step 1: Objectives of the Study

- Where is the poor population located mostly?

## ② Step 2: Methodology

- **Unsupervised:**  $K$ -means Clustering
  - I clustered households based on their financial strength.
- **Supervised:** Bagging, Random Forest, and Gradient Boosting
  - After  $K$ -means clustering, I modeled the clusters, using supervised learning, to understand the accuracy of the clustering.

# Workflow

## ④ Step 1: Objectives of the Study

- Where is the poor population located mostly?

## ② Step 2: Methodology

- **Unsupervised:** *K*-means Clustering
  - I clustered households based on their financial strength.
- **Supervised:** Bagging, Random Forest, and Gradient Boosting
  - After *K*-means clustering, I modeled the clusters, using supervised learning, to understand the accuracy of the clustering.

# Unsupervised Learning

## • Clustering

- Let  $C_k$  be the generic cluster of observations with  $k = 1, \dots, K$  which satisfies the two properties:
  - $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$ , where each observation belongs to at least one cluster.
  - $C_k \cap C_{k'} = \emptyset$  for all  $k \neq k'$ . Each observation belongs only to one cluster.
- $K$ -means cluster sets the *within – cluster variance* as small as possible, and it is defined as

$$\min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K W(C_k) \right\} = \min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\} \quad (5)$$

where  $|C_k|$  is the number of observations in the  $k$ -th cluster and  $W(C_k)$  is a measure of between distance.

# Unsupervised Learning

## • Clustering

- Let  $C_k$  be the generic cluster of observations with  $k = 1, \dots, K$  which satisfies the two properties:
  - $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$ , where each observation belongs to at least one cluster.
  - $C_k \cap C_{k'} = \emptyset$  for all  $k \neq k'$ . Each observation belongs only to one cluster.
- $K$ -means cluster sets the *within – cluster variance* as small as possible, and it is defined as

$$\min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K W(C_k) \right\} = \min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\} \quad (5)$$

where  $|C_k|$  is the number of observations in the  $k$ -th cluster and  $W(C_k)$  is a measure of between distance.

# Unsupervised Learning: K-Means Clustering

- I clustered one of the asset variables NW-HE<sup>7</sup>.
- When the number of cluster is  $K = 6$  the predictive powers of the supervised learning are the highest comparing to  $K! = 6$ .

|     | $K = 3$ | $K = 4$ | $K = 5$ | $K = 6$ | $K = 7$ | $K = 8$ | $K = 9$ | $K = 10$ |
|-----|---------|---------|---------|---------|---------|---------|---------|----------|
| MLM | 69.00   | 65.37   | 70.68   | 71.30   | 62.32   | 62.28   | 56.70   | 53.25    |
| BAG | 49.92   | 57.24   | 63.26   | 68.94   | 49.62   | 54.22   | 57.62   | 56.66    |
| RF  | 43.03   | 64.16   | 55.23   | 61.52   | 53.39   | 63.98   | 49.97   | 55.41    |

Note: Results were GBM were not converged for all the clusters, so I drop it.

|              | Cluster 1    | Cluster 2 | Cluster 3    | Cluster 4    | Cluster 5    | Cluster 6 |
|--------------|--------------|-----------|--------------|--------------|--------------|-----------|
| Income       | 1.092860e+06 | 85273.24  | 1.462134e+05 | 3.633347e+05 | 2.318359e+05 | 41603.10  |
| Observations | 117          | 6875      | 2387         | 284          | 827          | 22275     |

Note: Income is in Chinese renminbi (CNY). Based on the financial health of the households', the clusters were colored from green, yellow, orange, and red.

<sup>7</sup>Note that: the 3 asset owning variables, namely; Networth, NW-HE, and Liquid Assets are highly correlated.



# Unsupervised Learning: K-Means Clustering

- I clustered one of the asset variables NW-HE<sup>7</sup>.
- When the number of cluster is  $K = 6$  the predictive powers of the supervised learning are the highest comparing to  $K! = 6$ .

|     | $K = 3$ | $K = 4$ | $K = 5$ | $K = 6$ | $K = 7$ | $K = 8$ | $K = 9$ | $K = 10$ |
|-----|---------|---------|---------|---------|---------|---------|---------|----------|
| MLM | 69.00   | 65.37   | 70.68   | 71.30   | 62.32   | 62.28   | 56.70   | 53.25    |
| BAG | 49.92   | 57.24   | 63.26   | 68.94   | 49.62   | 54.22   | 57.62   | 56.66    |
| RF  | 43.03   | 64.16   | 55.23   | 61.52   | 53.39   | 63.98   | 49.97   | 55.41    |

Note: Results were GBM were not converged for all the clusters, so I drop it.

|              | Cluster 1    | Cluster 2 | Cluster 3    | Cluster 4    | Cluster 5    | Cluster 6 |
|--------------|--------------|-----------|--------------|--------------|--------------|-----------|
| Income       | 1.092860e+06 | 85273.24  | 1.462134e+05 | 3.633347e+05 | 2.318359e+05 | 41603.10  |
| Observations | 117          | 6875      | 2387         | 284          | 827          | 22275     |

Note: Income is in Chinese renminbi (CNY). Based on the financial health of the households', the clusters were colored from green, yellow, orange, and red.

<sup>7</sup>Note that: the 3 asset owning variables, namely; Networth, NW-HE, and Liquid Assets are highly correlated.

# Unsupervised Learning: K-Means Clustering

- I clustered one of the asset variables NW-HE<sup>7</sup>.
- When the number of cluster is  $K = 6$  the predictive powers of the supervised learning are the highest comparing to  $K! = 6$ .

|     | $K = 3$ | $K = 4$ | $K = 5$ | $K = 6$ | $K = 7$ | $K = 8$ | $K = 9$ | $K = 10$ |
|-----|---------|---------|---------|---------|---------|---------|---------|----------|
| MLM | 69.00   | 65.37   | 70.68   | 71.30   | 62.32   | 62.28   | 56.70   | 53.25    |
| BAG | 49.92   | 57.24   | 63.26   | 68.94   | 49.62   | 54.22   | 57.62   | 56.66    |
| RF  | 43.03   | 64.16   | 55.23   | 61.52   | 53.39   | 63.98   | 49.97   | 55.41    |

Note: Results were GBM were not converged for all the clusters, so I drop it.

|              | Cluster 1    | Cluster 2 | Cluster 3    | Cluster 4    | Cluster 5    | Cluster 6 |
|--------------|--------------|-----------|--------------|--------------|--------------|-----------|
| Income       | 1.092860e+06 | 85273.24  | 1.462134e+05 | 3.633347e+05 | 2.318359e+05 | 41603.10  |
| Observations | 117          | 6875      | 2387         | 284          | 827          | 22275     |

Note: Income is in Chinese renminbi (CNY). Based on the financial health of the households', the clusters were colored from green, yellow, orange, and red.

<sup>7</sup>Note that: the 3 asset owning variables, namely; Networth, NW-HE, and Liquid Assets are highly correlated.



# Micro App

- **Micro Shinyapp**<sup>8</sup> provides an interactive user interface, using the 2015 CHFS data-set.
- Micro Shinyapp is to map the all the households based on their Cluster where Cluster 2 and Cluster 6 can be interpreted as those with the lowest financial health.
- Map of the households can be found under the  Map tab and  Table tab prints out the pivot table.

---

<sup>8</sup>The link may fail to connect with the survey because of its size: <https://seymakalay87.shinyapps.io/micro/>



# Micro App

- **Micro Shinyapp**<sup>8</sup> provides an interactive user interface, using the 2015 CHFS data-set.
- Micro Shinyapp is to map the all the households based on their Cluster where Cluster 2 and Cluster 6 can be interpreted as those with the lowest financial health.
- Map of the households can be found under the  Map tab and  Table tab prints out the pivot table.

---

<sup>8</sup>The link may fail to connect with the survey because of its size: <https://seymakalay87.shinyapps.io/micro/>

# Micro App

- **Micro Shinyapp**<sup>8</sup> provides an interactive user interface, using the 2015 CHFS data-set.
- Micro Shinyapp is to map the all the households based on their Cluster where Cluster 2 and Cluster 6 can be interpreted as those with the lowest financial health.
- Map of the households can be found under the  Map tab and  Table tab prints out the pivot table.

---

<sup>8</sup>The link may fail to connect with the survey because of its size: <https://seymakalay87.shinyapps.io/micro/>

# Conclusion

- ❶ Implementing unsupervised for clustering households based on their financial health.
- ❷ Based on the sample data set, I mapped advantaged/disadvantaged households.
- ❸ Creating an interactive map for visualization which may help to policy makers to design suitable programs, aiming the most poor areas (e.g., Cluster 6 and Cluster 2) primarily.

# Conclusion

- ❶ Implementing unsupervised for clustering households based on their financial health.
- ❷ Based on the sample data set, I mapped advantaged/disadvantaged households.
- ❸ Creating an interactive map for visualization which may help to policy makers to design suitable programs, aiming the most poor areas (e.g., Cluster 6 and Cluster 2) primarily.

# Conclusion

- ❶ Implementing unsupervised for clustering households based on their financial health.
- ❷ Based on the sample data set, I mapped advantaged/disadvantaged households.
- ❸ Creating an interactive map for visualization which may help to policy makers to design suitable programs, aiming the most poor areas (e.g., Cluster 6 and Cluster 2) primarily.



# Outline

- 1 Chapter 1
  - Introduction
- 2 Chapter 2
  - Determinants of Access to Finance: A Bibliometric Literature Review
- 3 Chapter 3
  - Access to Credit: The Self-Employment Case in the Chinese Labor Market
- 4 Chapter 4
  - Predicting Financial Health of the Households Using Machine Learning Algorithms
- 5 Chapter 5
  - Conclusions

# Recap

## ④ Chapter 1: Determinants of Access to Finance: A Bibliometric Literature Review

- Finding influential aspects of the literature under "determinants of finance".
  - 210 published English articles coupled with content analysis.
  - Influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
  - Network analysis (Co-citation, Co-author, and Co-word).
  - Two main research streams (i) lending to small borrowers (ii) lending to big borrowers.
  - 13 future research questions.
  - A Shinyapp for the reproducible future studies: <https://seymakalay87.shinyapps.io/biblio/>.

## ⑤ Chapter 2: Access to Credit: The Self-Employment Case in the Chinese Labor Market

- Identifying the characteristics of the Chinese households to access to credit.
  - In total, 120 models, using 3 different asset variables interchangeable.
  - CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
  - A CRAN - Package for reproducibility: <https://seymakalay.github.io/pomodoro/index.html>.

## ⑧ Chapter 3: Predicting Financial Health of the Households Using ML Algorithms

- Locating advantaged and disadvantaged groups.
  - Implementing unsupervised and supervised analysis.
  - Finding the location of the advantaged/disadvantaged Chinese Households.
  - A Shinyapp for visualization: <https://seymakalay87.shinyapps.io/micro/>.

# Recap

## ④ Chapter 1: Determinants of Access to Finance: A Bibliometric Literature Review

- Finding influential aspects of the literature under "determinants of finance".
  - 210 published English articles coupled with content analysis.
  - Influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
  - Network analysis (Co-citation, Co-author, and Co-word).
  - Two main research streams (i) lending to small borrowers (ii) lending to big borrowers.
  - 13 future research questions.
  - A Shinyapp for the reproducible future studies: <https://seymakalay87.shinyapps.io/biblio/>.

## ② Chapter 2: Access to Credit: The Self-Employment Case in the Chinese Labor Market

- Identifying the characteristics of the Chinese households to access to credit.
  - In total, 120 models, using 3 different asset variables interchangeable.
  - CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
  - A CRAN - Package for reproducibility: <https://seymakalay.github.io/pomodoro/index.html>.

## ③ Chapter 3: Predicting Financial Health of the Households Using ML Algorithms

- Locating advantaged and disadvantaged groups.
  - Implementing unsupervised and supervised analysis.
  - Finding the location of the advantaged/disadvantaged Chinese Households.
  - A Shinyapp for visualization: <https://seymakalay87.shinyapps.io/micro/>.

# Recap

## ④ Chapter 1: Determinants of Access to Finance: A Bibliometric Literature Review

- Finding influential aspects of the literature under "determinants of finance".
  - 210 published English articles coupled with content analysis.
  - Influential aspects of the research stream (e.g., countries, affiliations, journals, authors, and articles).
  - Network analysis (Co-citation, Co-author, and Co-word).
  - Two main research streams (i) lending to small borrowers (ii) lending to big borrowers.
  - 13 future research questions.
  - A Shinyapp for the reproducible future studies: <https://seymakalay87.shinyapps.io/biblio/>.

## ② Chapter 2: Access to Credit: The Self-Employment Case in the Chinese Labor Market

- Identifying the characteristics of the Chinese households to access to credit.
  - In total, 120 models, using 3 different asset variables interchangeable.
  - CCP.0 & CCP.1 data split has higher predictive power and performs better to explain the Chinese household characteristics, for both  $y = \text{Access Loan}$  and  $y = \text{Loan Type}$ .
  - A CRAN - Package for reproducibility: <https://seymakalay.github.io/pomodoro/index.html>.

## ⑧ Chapter 3: Predicting Financial Health of the Households Using ML Algorithms

- Locating advantaged and disadvantaged groups.
  - Implementing unsupervised and supervised analysis.
  - Finding the location of the advantaged/disadvantaged Chinese Households.
  - A Shinyapp for visualization: <https://seymakalay87.shinyapps.io/micro/>.

Thank you

## Appendix - Chapter 2

|                       | Approaches                                     |   |  |
|-----------------------|--|---|--|
|                       | Systematic                                     | Semi-systematic<br>(Meta-analysis)  | Integrate  |
| (1) Purpose           | Synthesize and compare evidence                | Overview research area and track development over time  | Critique and synthesize                                    |
| (2) Research question | Specific                                       | Broad   | Narrow or broad  |
| (3) Research strategy | Systematic                                     | May or may not be systematic  | Usually not systematic                                     |
| (4) Characteristics   | Quantitative articles                          | Research articles   | Research articles, books, and other published documents    |
| (5) Analysis          | Quantitative                                   | Qualitative/Quantitative  | Qualitative  |
| (6) Contribution      | Evidence of effect, Inform policy and practice | State of knowledge, Themes in literature, Historical overview, Research agenda, Theoretical model | Taxonomy or classification, Theoretical model or framework |

Source at: Literature review as a research methodology: An overview and guidelines.

Table: Literature review approaches.

## Appendix - Chapter 2

|    | Country        | Article.No. | %Freq | SCP | %SCP  | MCP | %MCP  | TGC  | TGC/t  |
|----|----------------|-------------|-------|-----|-------|-----|-------|------|--------|
| 1  | USA            | 83          | 41.29 | 64  | 47.76 | 19  | 28.36 | 8120 | 97.80  |
| 2  | United Kingdom | 26          | 12.93 | 19  | 14.18 | 7   | 10.45 | 1024 | 39.40  |
| 3  | China          | 14          | 6.97  | 7   | 5.22  | 7   | 10.45 | 421  | 30.10  |
| 4  | Italy          | 7           | 3.48  | 5   | 3.73  | 2   | 2.99  | 125  | 17.90  |
| 5  | India          | 6           | 2.99  | 5   | 3.73  | 1   | 1.49  | 354  | 59.00  |
| 6  | France         | 5           | 2.49  | 2   | 1.49  | 3   | 4.48  | 206  | 41.20  |
| 7  | Germany        | 5           | 2.49  | 2   | 1.49  | 3   | 4.48  | 491  | 98.20  |
| 8  | Netherlands    | 4           | 1.99  | 2   | 1.49  | 2   | 2.99  | 502  | 125.50 |
| 9  | New Zealand    | 4           | 1.99  | 4   | 2.99  | 0   | 0.00  | 142  | 35.50  |
| 10 | South Africa   | 4           | 1.99  | 4   | 2.99  | 0   | 0.00  | 32   | 8.00   |

Note: The table is sorted based on total number of Article.No. %Freq, %SCP, and %MCP are the percentage of the total Article.No., SCP, and MCP, respectively.

|    | Affiliation                     | Articles | TLC | TLC/t | TGC  | TGC/t  |
|----|---------------------------------|----------|-----|-------|------|--------|
| 1  | World Bank                      | 14       | 69  | 6.24  | 2558 | 197.35 |
| 2  | Georgetown Univ                 | 2        | 53  | 2.55  | 297  | 14.62  |
| 3  | Dartmouth Coll                  | 1        | 31  | 1.72  | 237  | 13.17  |
| 4  | Wellesley Coll                  | 1        | 31  | 1.72  | 237  | 13.17  |
| 5  | Tilburg Univ                    | 5        | 30  | 3.13  | 634  | 55.68  |
| 6  | Fed Reserve Syst                | 2        | 28  | 1.48  | 239  | 12.91  |
| 7  | Harvard Univ                    | 5        | 28  | 2.58  | 1963 | 144.34 |
| 8  | Ctr Naval Anal                  | 1        | 26  | 1.13  | 111  | 4.83   |
| 9  | German Inst Econ Res Diw Berlin | 2        | 22  | 2.34  | 195  | 28.75  |
| 10 | Robert Gordon Univ              | 2        | 22  | 1.89  | 177  | 15.64  |

Note: The table is sorted based on TLC.

## Appendix - Chapter 2

|    | Journal                              | Article.No. | TLC | TLC/t | TGC  | TGC/t |
|----|--------------------------------------|-------------|-----|-------|------|-------|
| 1  | World Development                    | 14          | 50  | 6.49  | 683  | 88.75 |
| 2  | Journal Of Development Studies       | 10          | 18  | 2.29  | 200  | 23.32 |
| 3  | Journal Of Banking & Finance         | 8           | 23  | 2.28  | 1220 | 96.72 |
| 4  | Small Business Economics             | 7           | 18  | 2.39  | 230  | 29.86 |
| 5  | Environment And Planning A           | 5           | 1   | 0.07  | 194  | 11.72 |
| 6  | Journal Of International Development | 5           | 5   |       | 188  |       |
| 7  | Sustainability                       | 5           | 4   | 2.50  | 46   | 25.83 |
| 8  | Emerging Markets Finance And Trade   | 4           | 6   | 1.29  | 68   | 10.49 |
| 9  | Entrepreneurship Theory And Practice | 4           | 15  | 1.15  | 729  | 56.38 |
| 10 | Finance Research Letters             | 4           | 2   | 0.53  | 76   | 29.07 |
|    | Journal                              | Article.No. | TLC | TLC/t | TGC  | TGC/t |
| 1  | Journal Of Banking & Finance         | 8           | 23  | 2.28  | 1220 | 96.72 |
| 2  | World Development                    | 14          | 50  | 6.49  | 683  | 88.75 |
| 3  | Journal Of Financial Economics       | 2           | 16  | 1.23  | 917  | 70.54 |
| 4  | Management Science                   | 1           | 2   | 0.29  | 439  | 62.71 |
| 5  | Entrepreneurship Theory And Practice | 4           | 15  | 1.15  | 729  | 56.38 |
| 6  | Quarterly Journal Of Economics       | 1           | 14  | 0.88  | 866  | 54.12 |
| 7  | Annual Review Of Sociology           | 1           | 4   | 0.31  | 689  | 53.00 |
| 8  | Journal Of Finance                   | 2           | 5   | 0.22  | 812  | 52.86 |
| 9  | Small Business Economics             | 7           | 18  | 2.39  | 230  | 29.86 |
| 10 | Finance Research Letters             | 4           | 2   | 0.53  | 76   | 29.07 |

Note: The table is sorted by Article.No. (top) and TGC/t (bottom).



## Appendix - Chapter 2

|    | 1st Author    | Affiliation          | Article.No. | TLC | TLC/t | TGC  | TGC/t |
|----|---------------|----------------------|-------------|-----|-------|------|-------|
| 1  | Beck T        | World Bank           | 5           | 21  | 1.79  | 1111 | 83.23 |
| 2  | Wylly Ek      | Rutgers State Univ   | 3           | 4   | 0.23  | 148  | 8.97  |
| 3  | Agier I       | Univ Libre Bruxelles | 2           | 7   | 0.87  | 96   | 12.00 |
| 4  | Allen F       | Imperial Coll London | 2           | 14  | 2.74  | 175  | 31.74 |
| 5  | Asiedu E      | Univ Kansas          | 2           | 16  | 1.90  | 81   | 9.61  |
| 6  | Bates T       | Wayne State Univ     | 2           | 1   | 0.04  | 38   | 3.80  |
| 7  | Bayer P       | Duke Univ            | 2           | 0   | 0.00  | 25   | 7.58  |
| 8  | Black Ha      | Univ Tennessee       | 2           | 6   | 0.33  | 26   | 1.43  |
| 9  | Carter S      | Univ Sterling        | 2           | 9   | 0.83  | 283  | 29.26 |
| 10 | Cavalluzzo Ks | Georgetown Univ      | 2           | 53  | 2.55  | 297  | 14.62 |
|    | 1st Author    | Affiliation          | Article.No. | TLC | TLC/t | TGC  | TGC/t |
| 1  | Beck T        | World Bank           | 5           | 21  | 1.79  | 1111 | 83.23 |
| 2  | Fernandes D   | Erasmus Univ         | 1           | 2   | 0.29  | 439  | 62.71 |
| 3  | Khwaja Ai     | Harvard Univ         | 1           | 14  | 0.88  | 866  | 54.12 |
| 4  | Claessens S   | Int Monetary Fund    | 2           | 16  | 1.21  | 711  | 53.85 |
| 5  | Pager D       | Princeton Univ       | 1           | 4   | 0.31  | 689  | 53.00 |
| 6  | Campbell Jy   | Harvard Univ         | 1           | 1   | 0.07  | 769  | 51.27 |
| 7  | Allen F       | Imperial Coll London | 2           | 14  | 2.74  | 175  | 31.74 |
| 8  | Carter S      | Univ Sterling        | 2           | 9   | 0.83  | 283  | 29.26 |
| 9  | Hastings Js   | Brown Univ           | 1           | 2   | 0.25  | 198  | 24.75 |
| 10 | Houston Jf    | Univ Florida         | 1           | 1   | 0.14  | 172  | 24.57 |

Note: The table is sorted by sorted based on Article.No. (top) and TGC/t (bottom).

## Appendix - Chapter 2



**Figure:** The relationship between authors in "determinants of finance" through bibliometric co-authorship (collaboration) analysis between 1985 and 2020.

## Appendix - Chapter 3

| Variable                     | Definition  |
|------------------------------|---|
| <u>independent variables</u> |   |
| $x_1$ Gender                 | $Sex.1 = male = 1, Sex.0 = female = 0$  |
| $x_2$ Marital Status         | $married = 1, otherwise = 0$  |
| $x_3$ Age                    | household's age, in years.  |
| $x_4$ Employed               | $employed = 1, otherwise = 0$   |
| $x_5$ Education              | $Educ.1 = high\ school\ or\ higher = 1, Educ.0 = otherwise = 0$   |
| $x_6$ Party                  | $CCP.1 = affiliation\ with\ Chinese\ Communist\ Party\ (CCP) = 1, CCP.0 = otherwise = 0$  |
| $x_7$ HR                     | $Urb = urban = 1, Rrl = rural = 0$  |
| $x_8$ Region                 | west, east, and center.   |
| $x_9$ Fin.Inter              | $Fin.Inter = 1$ if the household head house is in 1 km range to formal institution, otherwise $Fin.Inter = 0$   |
| $x_{10}$ Fin.Knowledge       | $Fin.Knowledge = 1$ if the household head has a finance class or defined him/herself having a well financial knowledge, otherwise $Fin.Knowledge = 0$   |
| $x_{11}$ Income              | household's income, in CNY.   |
| $x_{12}$ Net-worth           | The value of financial and non-financial assets minus liabilities, in CNY.  |
| $x_{13}$ NW-HE               | Net-worth minus home equity, in CNY.  |
| $x_{14}$ Liquid Assets       | Cash and other easily cash-able assets, in CNY.   |
| <u>dependent variables</u>   |   |
| $y_1$ Access to loan         | $Access\ to\ loan = 1$ if the household head has any type of loan (e.g formal, informal, and/or both); otherwise $Access\ to\ loan = 0$ .   |
| $y_2$ Access to loan type    | if the household head has formal, informal, or both loans $Access\ to\ loan\ type$ is equal to 1, 2, 3, respectively; otherwise $Access\ to\ loan\ type = 0$ , which indicates the household head does not have any type of loan. |

Note: HR stands for Household Registration. In equations  $x_{12}$ ,  $x_{13}$ , and  $x_{14}$  were use interchangeability.

Table: Definitions of the independent variables.

## Appendix - Chapter 3

| Abbreviation         | Definition  |
|----------------------|---|
| <u>data splits</u>   |   |
| Urb & Rrl            | CHFS data-set was split into urban and rural.   |
| Educ.0 & Educ.1      | CHFS data-set was split into Educ.0 and Educ.1.   |
| CCP.0 & CCP.1        | CHFS data-set was split into CCP.0 and CCP.1  |
| Sex.0 & Sex.1        | CHFS data-set was split into Sex.0 and Sex.1  |
| <u>linear models</u> |   |
| GLM                  | To model access to credit. $y = 1$ If the household's head has any type of loan (e.g. Formal, Informal, or Both), otherwise $y = 0$ |
| MLM                  | To model access to loan type. $y = 1, 2, 3,$ or $4$ If the household's head has Formal, Informal, Both, or No.loan, respectively.   |
| <u>ml models</u>     |   |
| BAG                  | Bag tree  |
| RF                   | Random forest   |
| GBM                  | Gradient boosting   |
| <u>loan types</u>    |   |
| Fml                  | If household's head has only Formal loan.   |
| Infml                | If household's head has only Informal loan.   |
| Both                 | If household's head has both Formal and Informal loan.  |

Note: Urb & Rrl, Educ.0 & Educ.1, CCP.0 & CCP.1, and Sex.0 & Sex.1 are defined in Table 3.

Table: Definitions of abbreviations.

# Appendix - Chapter 4

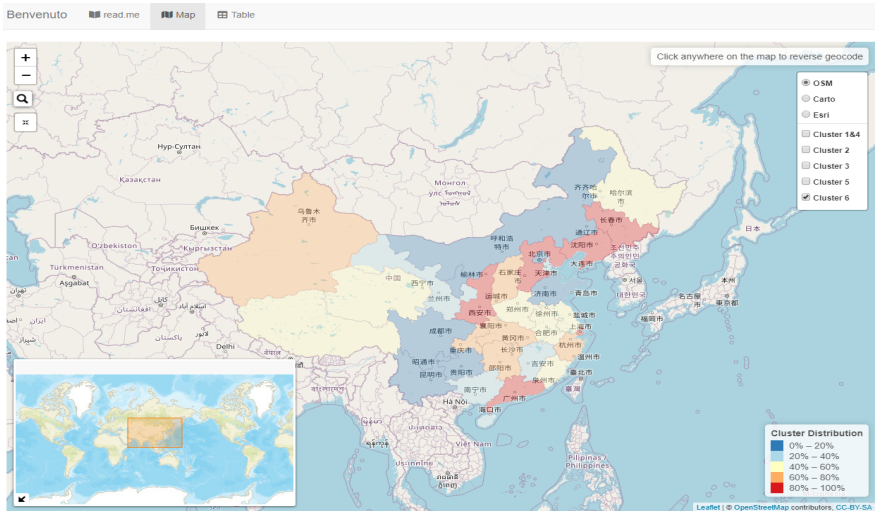


Figure: Map